Composite Neyman-Pearson Hypothesis Testing with a Known Hypothesis

Parham Boroumand
University of Cambridge
pb702@cam.ac.uk

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—We propose a composite hypothesis test in the Neyman-Pearson setting where the null distribution is known and the alternative distribution belongs to a certain family of distributions. The proposed test interpolates between Hoeffding's test and the likelihood ratio test and achieves the optimal error exponent tradeoff for every distribution in the family. In addition, the proposed test is shown to attain the type-I error probability prefactor of $n^{\frac{\bar{d}-1}{2}}$, where $\bar{d}$ is the dimension of the family of distributions projected onto a relative entropy ball centered at the null distribution. This can be significantly smaller than the prefactor $n^{\frac{d-2}{2}}$ achieved by the Hoeffding's test where d is the dimension of the probability simplex. In addition, the proposed test achieves the optimal type-II error probability prefactor for every distribution in the family.**

I. Introduction and Preliminaries

Consider the following composite binary hypothesis testing problem where an observation $\boldsymbol{x} = (x_1, \dots, x_n)$ is generated in an i.i.d. fashion from either of two possible distributions P_0 or P_1 defined on a probability simplex $\mathcal{P}(\mathcal{X})$ with alphabet size $|\mathcal{X}| < \infty$. The type of $\boldsymbol{x}$ is defined as $\hat{T}_{\boldsymbol{x}}(a) = \frac{N(a|\boldsymbol{x})}{n}$, where $N(a|\boldsymbol{x})$ is the number of occurrences of symbol $a \in \mathcal{X}$ in sequence $\boldsymbol{x}$. The set of all sequences of length n with type P, denoted by $\mathcal{T}_P^n$. The set of types formed with length n sequences on the simplex $\mathcal{P}(\mathcal{X})$ is denoted as $\mathcal{P}_n(\mathcal{X})$. We assume that the first hypothesis is known and that distribution P_1 belongs to a known family of distributions denoted by $\mathcal{P}_1$ and characterized by

$$\mathcal{P}_1 = \big\{P_1(\theta) : \theta \in \Theta, \Theta \subseteq \mathbb{R}^{\tilde{d}}\big\}, \tag{1}$$

where $\tilde{d} \le d$, and $d = |\mathcal{X}| - 1$, which is the dimension of probability simplex in $\mathbb{R}^{d+1}$. Therefore, P_1 is uniquely characterized by the choice of parameter θ. Also, assume that Θ is a compact set and $P_1(\theta)$ be a continuous function of θ. Let $\phi : \mathcal{X}^n \to \{0, 1\}$ be a hypothesis test that decides which distribution generated the observation $\boldsymbol{x}$. We also assume that both $P_0(x) > 0, P_1(x) > 0$ for each $x \in \mathcal{X}$. We consider deterministic tests ϕ that decide in favor of P_0 if $\boldsymbol{x} \in \mathcal{A}_0$, where $\mathcal{A}_0 \subset \mathcal{X}^n$ is the decision region for the first hypothesis. We define $\mathcal{A}_1 = \mathcal{X}^n \setminus \mathcal{A}_0$ to be the decision region for the second hypothesis. The two possible pairwise error probabilities measure the test performance. The type-I and type-II error probabilities are defined as

$$\epsilon_0(\phi) = \sum_{\boldsymbol{x} \in \mathcal{A}_1} P_0^n(\boldsymbol{x}), \quad \epsilon_1(\phi|P_1(\theta)) = \sum_{\boldsymbol{x} \in \mathcal{A}_0} P_1^n(\theta)(\boldsymbol{x}). \tag{2}$$

The optimal error exponent tradeoff (E_0, E_1) is defined as

$$E_1^*(E_0, \theta) \triangleq \sup \big\{E_1 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \\ \epsilon_0(\phi) \le e^{-nE_0} \quad \text{and} \quad \epsilon_1(\phi|P_1(\theta)) \le e^{-nE_1}\big\}. \tag{3}$$

Assuming $P_1(\theta)$ is known, the likelihood ratio test defined as

$$\phi^{\text{lrt}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}} \| P_0) - D(\hat{T}_{\boldsymbol{x}} \| P_1(\theta)) \ge \gamma\big\}. \tag{4}$$

achieves the optimal error exponent trade-off in (3) [1], [2]. For composite hypothesis testing, Hoeffding proposed the following generalized likelihood ratio test [3]

$$\phi^{\text{glrt}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}} \| P_0) > E_0\big\}. \tag{5}$$

It is known that Hoeffding's test can achieve the best error exponent trade-off in (3) for any family of distributions $\mathcal{P}_1$ [3], [4]. By Sanov's theorem [5], the error exponent of Hoeffding's test is given by

$$E_0(\phi^{\text{glrt}}) = E_0, \tag{6}$$

$$E_1(\phi^{\text{glrt}}) = \min_{\substack{Q \in \mathcal{P}(\mathcal{X}), \\ D(Q\|P_0) \le E_0}} D(Q \| P_1), \tag{7}$$

where the minimizing distribution in (7) is given by

$$Q_\mu(x) = \frac{P_0^{\frac{\mu}{1+\mu}}(x) P_1^{\frac{1}{1+\mu}}(x)}{\sum_{a \in \mathcal{X}} P_0^{\frac{\mu}{1+\mu}}(a) P_1^{\frac{1}{1+\mu}}(a)}, \quad \mu \ge 0, \tag{8}$$

and the μ is the solution to $D(Q_\mu \| P_0) = E_0$ [2]. Using large deviations refinement [6]–[8], the type-I error probability of the likelihood ratio test is

$$\epsilon_0(\phi^{\text{lrt}}) = \frac{1}{\sqrt{n}} e^{-nE_0}\big(c + o(1)\big), \tag{9}$$

while, Hoeffding's test type-I error probability is given by [7], [9]

$$\epsilon_0(\phi^{\text{glrt}}) = n^{\frac{|\mathcal{X}|-3}{2}} e^{-nE_0}\big(c' + o(1)\big), \tag{10}$$

This work was supported in part by the European Research Council under Grant 725411, and by the Spanish Ministry of Economy and Competitiveness under Grant PID2020-116683GB-C22.

where c, c' are constants that only depend on P_0, P_1, and the corresponding test thresholds. As a result, when the alphabet size is large, Hoeffding's test, although attaining the optimal error exponent tradeoff, suffers in prefactor when compared to the likelihood ratio's $\frac{1}{\sqrt{n}}$ for observation alphabets such that $|\mathcal{X}| > 2$. In this paper, we propose a test which exploits the geometry of the family of alternative distributions $\mathcal{P}_1$ to reduce this gap between likelihood ratio test and Hoeffding's test pre-factors.

II. Main Results

We show that if there is some low dimensional structure on the family $\mathcal{P}_1$, the prefactor can be controlled by the dimension of the family $\tilde{d}$ instead of the dimension of probability simplex d. Our proposed test is the intersection of the decision regions $\mathcal{A}_0$ of all likelihood ratio test between P_0 and every distribution P_1 in the family $\mathcal{P}_1(\theta)$. To introduce the main idea, we give the following example.

Example 1. Assume that the first hypothesis is distributed by P_0 and the family of the distributions $\mathcal{P}_1$ is the union of two exponential families generated by P_0, P_1 and P_0, P_1', i.e.,

$$\begin{aligned}\mathcal{P}_1 = &\left\{P : P = \frac{P_0^{\lambda}(x)P_1^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_0^{\lambda}(a)P_1^{1-\lambda}(a)}, \lambda \subset [0,1]\right\} \\ &\cup \left\{P : P = \frac{P_0^{\lambda}(x)P_1'^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_0^{\lambda}(a)P_1'^{1-\lambda}(a)}, \lambda' \subset [0,1]\right\}. \quad (11)\end{aligned}$$

In this case, one can use a test combining the result of two likelihood ratio tests and achieves the optimal prefactor $\frac{1}{\sqrt{n}}$. Let the test be

$$\begin{aligned}\phi(\hat{T}_{\boldsymbol{x}}) =& \mathbb{1}\Big\{D(\hat{T}_{\boldsymbol{x}}\|P_0) - D(\hat{T}_{\boldsymbol{x}}\|P_1) > \gamma \quad \text{or} \\ & D(\hat{T}_{\boldsymbol{x}}\|P_0) - D(\hat{T}_{\boldsymbol{x}}\|P_1') > \gamma'\Big\}, \qquad (12)\end{aligned}$$

where γ, γ' are chosen such that for each likelihood ratio test, the error exponent under the first hypothesis is some fixed exponent E_0. This test is the intersection of two hyperplanes between P_0, P_1, and P_0, P_1' such that both hyperplanes are tangent to the Hoeffding's decision region. The optimality of the test in the prefactor is the direct result of the union bound. Figure 1 illustrates the decision region.

Let the relative entropy ball defined as

$$\mathcal{B}(E_0) = \{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) < E_0\}, \qquad (13)$$

and $\mathcal{P}_1^*$ to be the I-projection of $\mathcal{P}_1$ into the relative entropy ball $\mathcal{B}(E_0)$ [10], i.e.,

$$\mathcal{P}_1^* = \{P_1^*(\theta) : \theta \in \Theta\}, \qquad (14)$$

where

$$P_1^*(\theta) = \mathop{\arg\min}_{Q: D(Q\|P_0)\le E_0} D\big(Q\|P_1(\theta)\big). \qquad (15)$$

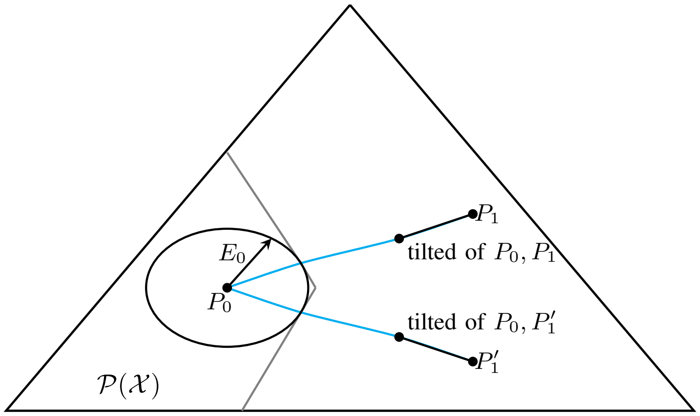

Fig. 1. Composite hypothesis testing with second hypothesis restricted to the union of two exponential families

Note that $P_1^*(\theta)$ is a continuous function of θ by the Berge's maximum theorem [11]. We propose the following hypothesis test

$$\phi(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\Big\{\exists\theta\in\Theta \text{ s.t. } D(\hat{T}_{\boldsymbol{x}}\|P_0) - D(\hat{T}_{\boldsymbol{x}}\|P_1^*(\theta)) \ge E_0\Big\}. \qquad (16)$$

The proposed test is the intersection of the likelihood ratio tests, such that the resulting hyperplane of the likelihood ratio test is tangent to the relative entropy ball of radius E_0 centered at P_0 for every $P_1(\theta)$. We note that the test proposed in [12] is different from the test in (16). The next theorem shows that the type-I error exponent of the proposed test equals E_0, i.e., it is independent of $P_1(\theta)$.

Theorem 1. *For every $P_1 \in \mathcal{P}_1$, the type-I error exponent of the proposed test ϕ in* (16) *satisfies*

$$\lim_{n\to\infty} -\frac{1}{n}\log\epsilon_0(\phi) = E_0. \qquad (17)$$

Proof. Using Sanov's Theorem [1] the type-I error exponent of the proposed test is given by

$$E_0(\phi) = \min_{Q\in\mathcal{Q}_0^c} D(Q\|P_0), \qquad (18)$$

where

$$\mathcal{Q}_0 = \bigcap_{\theta\in\Theta}\Big\{Q : D(Q\|P_0) < E_0 + D(Q\|P_1^*(\theta))\Big\}. \qquad (19)$$

From the definition, we get that $\mathcal{Q}_0 \subset \mathcal{B}(E_0)$ since for every $Q \in \mathcal{B}(E_0)$ we have $D(Q\|P_0) < E_0$, thus $Q \in \mathcal{Q}_0$. Therefore, we have that

$$\begin{aligned}E_0(\phi) &\ge \min_{Q\in\mathcal{B}(E_0)^c} D(Q\|P_0) \qquad (20)\\ &= E_0. \qquad (21)\end{aligned}$$

Moreover, for every θ, $P_1^*(\theta) \in \mathcal{Q}_0^c$ and $D(P_1^*(\theta)\|P_0) = E_0$, hence the type-I error exponent lower bound in (20) is achievable and the minimum in (18) equals to E_0. □

We next show that the type-II error probability of the proposed test achieves the optimal type-II error exponent and a prefactor that equals that of the likelihood ratio test for every $P_1(\theta) \in \mathcal{P}_1$ [6], [8].

Theorem 2. *For every $P_1(\theta)$, the type-II probability of error of the proposed test* (16) *satisfies*

$$\epsilon_1(\phi|P_1(\theta)) \leq \frac{1}{\sqrt{n}} e^{-nE_1^*(E_0,P_1(\theta))}(c(\theta)+o(1)), \qquad (22)$$

where

$$E_1^*(E_0,P_1(\theta)) = \min_{D(Q\|P_0)\leq E_0} D(Q\|P_1(\theta)) \qquad (23)$$

is the optimal type-II error exponent under the probability distribution $P_1(\theta)$ for the fixed type-I error exponent.

Proof. We first show that the test achieves the optimal exponent under each $P_1(\theta) \in \mathcal{P}_1$. For every $P_1(\theta)$, consider the following set

$$\mathcal{Q}_1(\theta) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1^*(\theta)) \geq E_0 \big\}. \qquad (24)$$

Then for every $\hat{T}_{\boldsymbol{x}} \in \mathcal{Q}_1(\theta)$ we have $\phi(\hat{T}_{\boldsymbol{x}}) = 1$. Hence, $\mathcal{Q}_1(\theta) \subseteq \mathcal{A}_1$, and by Sanov's theorem we have

$$E_1(E_0,P_1(\theta)) \geq \inf_{Q\in\mathcal{Q}_1^c(\theta)} D(Q\|P_1(\theta)). \qquad (25)$$

We show that the optimization in (25) equals to $E_1^*(E_0,P_1(\theta))$ as letting $Q = P_1^*(\theta)$ satisfy the KKT conditions, and since the optimization problem is convex, the KKT conditions are also sufficient. Writing the Lagrangian we obtain

$$\begin{aligned} L(Q,\lambda,\nu) =& D(Q\|P_1(\theta)) + \lambda\big(D(Q\|P_0) - D(Q\|P_1^*(\theta)) \\ & - E_0\big) + \nu\Big(\sum_{x\in\mathcal{X}} Q(x) - 1\Big). \end{aligned} \qquad (26)$$

Differentiating with respect to $Q(x)$ and setting to zero we have

$$1 + \log\frac{Q(x)}{P_1(\theta)(x)} + \lambda \log\frac{P_1^*(\theta)(x)}{P_0(x)} + \nu = 0. \qquad (27)$$

Solving equation (27) for every $x \in \mathcal{X}, Q \in \mathcal{P}(\mathcal{X})$ we obtain

$$Q_\lambda(x) = \frac{P_0^\lambda(x) P_1(\theta)(x) P_1^*(\theta)^{-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_0^\lambda(a) P_1(\theta)(a) P_1^*(\theta)^{-\lambda}(a)}. \qquad (28)$$

It can be shown that $P_1^*(\theta)$ is the tilted distribution of $P_0, P_1(\theta)$, therefore $Q_\lambda(x)$ will also be the tilted distribution of $P_0, P_1(\theta)$ [2]. Furthermore, by the complementary slackness condition [13] $D(Q_\lambda\|P_0) - D(Q_\lambda\|P_1^*(\theta)) = E_0$, hence the solution to the optimization is $Q_\lambda = P_1^*(\theta)$, and the infimum in (25) is equal to $E_1^*(E_0,P_1(\theta))$. Finally, by Theorem 1, for every θ the type-I error exponent is E_0, therefore the type-II error exponent cannot be larger than $E_1^*(E_0,P_1(\theta))$ as $E_1^*(E_0,P_1(\theta))$ is the optimal error exponent tradeoff in (3) for every $\theta \in \Theta$; hence (25) satisfies with equality.

Next, we show the prefactor decay relation in (22). As $\mathcal{Q}_1(\theta) \subseteq \mathcal{A}_1$ we have

$$\epsilon_1(\phi|P_1(\theta)) = \sum_{\boldsymbol{x}\in\mathcal{A}_1^c} P_1^n(\theta)(\boldsymbol{x}) \leq \sum_{\boldsymbol{x}\in\mathcal{Q}_1^c(\theta)} P_1^n(\theta)(\boldsymbol{x}). \qquad (29)$$

Furthermore, the decision region in (24) is equal to the likelihood ratio test between P_0 and $P_1^*(\theta)$. By [6], [8], the prefactor of likelihood ratio test is $\frac{1}{\sqrt{n}}$ for any pair of distributions P_0, P_1 hence we can further upper bound (29) to get (22). □

So far, we have shown that the proposed test achieves the optimal error exponent tradeoff. Also, for every θ, the type-II prefactor is $\frac{1}{\sqrt{n}}$ which is the same as if the parameter θ was known and the likelihood ratio test which achieves the optimal error tradeoff is used. We next show that the test in (16) achieves an improved prefactor with respect to Hoeffding's test. In particular, the power of n in the prefactor is a function of the dimension of $\mathcal{P}_1^*$; this can be much smaller than the dimension of the probability space.

Theorem 3. *Let $\mathcal{Q}_1 = \mathcal{Q}_0^c$ have a minimally smooth boundary. Then for every $P_1(\theta)$, the type-I error probability satisfies*

$$\epsilon_0(\phi) \leq n^{\frac{\bar{d}-1}{2}} e^{-nE_0}(c + o(1)), \qquad (30)$$

where $\bar{d}$ is the dimension of the manifold $\mathcal{P}_1^*$.

Example 2. We present a numerical example to illustrate the performance of the proposed hypothesis test in (16). Consider the distribution $P_0 = [0.25, 0.25, 0.25, 0.25]$, and let

$$\begin{aligned} \mathcal{P}_1 = \Bigg\{ & P_1 : P_1(x) = \frac{P_0^{1-\lambda}(x) P_1^{*\lambda}(x)}{\sum_{a\in\mathcal{X}} P_0^{1-\lambda}(a) P_1^{*\lambda}(a)}, \\ & \lambda \subset [1.2, 1.5], P_1^* \in \mathcal{P}_1^* \Bigg\}, \end{aligned} \qquad (31)$$

$$\begin{aligned} \mathcal{P}_1^* = \Bigg\{ & P_1^* : D(P_1^*\|P_0) = 0.001 \\ & \sum_{i=1}^4 P_1^*(i) = 1, P_1^*(1) + P_1^*(2) = 0.5 \Bigg\}. \end{aligned} \qquad (32)$$

It can be shown that $\mathcal{P}_1^*$ is the I-projection of $\mathcal{P}_1$ onto the relative entropy ball $\mathcal{B}(E_0 = 0.001)$, and that the boundary of the decision region is a smooth manifold. We also assume $P_1 = [0.3244, 0.1766, 0.1862, 0.3128] \in \mathcal{P}_1$ is generating the samples from the family $\mathcal{P}_1$. Each point in the figure is obtained by estimating the average error probability as follows. For each length of the test sequence n, we estimate the type-I and type-II error probabilities of the test in (16) as well as optimal likelihood ratio test assuming the knowledge of P_1, and the Hoeffding's generalized likelihood ratio test over 10^6 independent experiments. In Figure 2, we have plotted $\log\frac{\epsilon_0(\phi)}{(\sqrt{n})^{k(\phi)} e^{-nE_1}}$ where $k(\phi)$ equals to $-1, 0, 1$ for the likelihood ratio test, the proposed test, and Hoeffding's test respectively. It can be observed that all three curves converge to constants, hence the type-I error probability of the proposed test outperforms the Hoeffding's test as $\epsilon_0 \asymp e^{-nE_0}$ while $\epsilon_0^{\mathrm{H}} \asymp \sqrt{n} e^{-nE_0}$. In addition, in Figure 3 we plot $\log\frac{\epsilon_1(\phi)}{\frac{1}{\sqrt{n}} e^{-nE_1}}$ and we observe that the all three tests achieve the $\frac{1}{\sqrt{n}}$ prefactor, though with different constants.

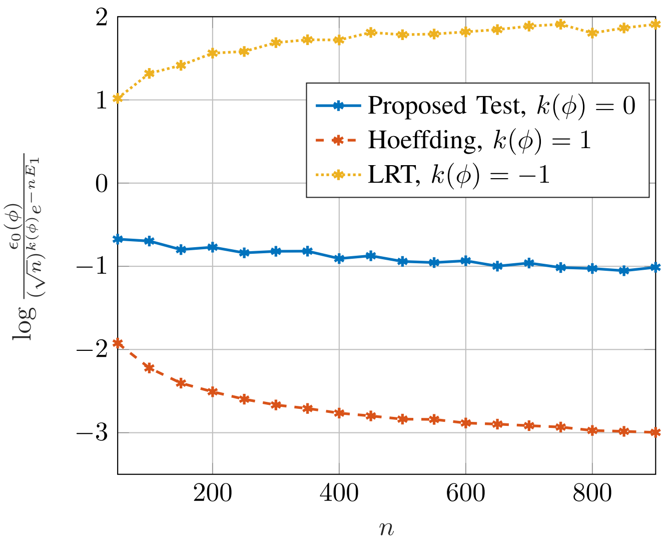

Fig. 2. Type-I error probabilities for the proposed test ($k(\phi) = 0$), likelihood ratio test ($k(\phi) = -1$), and Hoeffding's test ($k(\phi) = 1$).

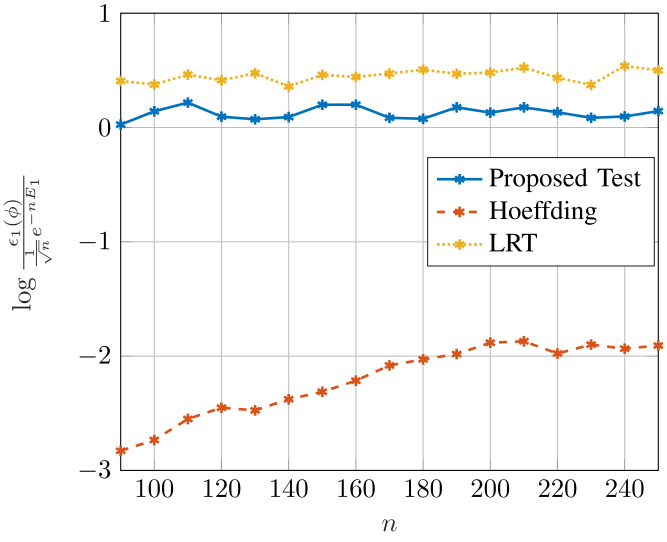

Fig. 3. Type-II error probabilities for the proposed test, likelihood ratio test, and Hoeffding's test.

Appendix A
Proof of Theorem 3

We find refined asymptotics of the type-I probability of error the proposed test (16). To find the polynomial decay of the error probability, we use the following theorem from [6], to approximate the summation by an integral and then use the saddlepoint approximation [14]–[16]. Also, we take similar steps to [7] to calculate the integral. We use the shorthand notation $a_n \asymp b_n$ for any two positive real sequences such that $\log \frac{a_n}{b_n} = \mathcal{O}(1)$.

Theorem 4. *[6] Suppose $\psi : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous function and the open set $\mathcal{D} \subseteq \mathbb{R}^d$ has minimally smooth boundary. Then*

$$\sum_{\boldsymbol{q} \in \mathcal{D} \cap \mathcal{P}_n(\mathcal{X})} e^{-n\psi(\boldsymbol{q})} \asymp n^d \int_{\mathcal{D}} e^{-n\psi(\boldsymbol{q})} \, d\boldsymbol{q}. \tag{33}$$

The exact definition of minimally smooth boundary can be found in [6]. In this paper, we give examples of sets with smooth boundaries and hence minimally smooth.

Using the method of types, the type-I probability of error can be written as

$$\begin{aligned} \epsilon_0(\phi) &= \sum_{\boldsymbol{x}:\phi(\hat{T}_{\boldsymbol{x}})=1} P_0(\boldsymbol{x}) && (34)\\ &= \sum_{Q \in \mathcal{Q}_1 \cap \mathcal{P}_n(\mathcal{X})} P_0(\mathcal{T}_Q^n) && (35)\\ &= \sum_{Q \in \mathcal{Q}_1 \cap \mathcal{P}_n(\mathcal{X})} \cdot n^{\frac{-|\mathcal{X}|+1}{2}} e^{-nD(Q\|P_0)} (c' + o(1)) && (36) \end{aligned}$$

where (35) is because the test is only depending on the type of the samples and in (36) we have used that [10]

$$P_0(\mathcal{T}_Q^n) \asymp n^{\frac{-|\mathcal{X}|+1}{2}} e^{-nD(Q\|P_0)}. \tag{37}$$

By the continuous differentiability of $\psi(\boldsymbol{q}) = D(\boldsymbol{q}\|P_0)$ in $\boldsymbol{q}$ and the minimally smoothness of the boundary of $\mathcal{Q}_1$ by the assumption, the conditions of the theorem 4 are satisfied and we can approximate the summation in (36) by the integral

$$\sum_{\boldsymbol{q} \in \mathcal{Q}_1} e^{-nD(\boldsymbol{q}\|P_0)} \asymp n^{|\mathcal{X}|-1} \int_{\boldsymbol{q} \in \mathcal{Q}_1} e^{-nD(\boldsymbol{q}\|P_0)} d\boldsymbol{q}, \tag{38}$$

where we restrict the space to the probability simplex by writing $P(|\mathcal{X}|) = 1 - \sum_{x=1}^{|\mathcal{X}|-1} P(x)$ for any probability distribution. To calculate the integral in the $d = |\mathcal{X}| - 1$ dimensional space over the set $\mathcal{Q}_1$, for every point $\boldsymbol{q}$ in $\mathcal{P}_1^*$ which is also in the decision region $\mathcal{Q}_1$, we define a local coordinate system centered at $\boldsymbol{q}$ such that dth coordinate w_d is parallel with

$$\boldsymbol{n}(\boldsymbol{q}) = \nabla_{\boldsymbol{q}} D(\boldsymbol{q}\|P_0), \tag{39}$$

the normal direction to the level surface $D(\boldsymbol{q}\|P_0) = E_0$. Furthermore, In this local coordinate system we denote the boundaries of $\mathcal{B}(E_0)$, $\mathcal{Q}_1$ by $w_d = f_{\boldsymbol{q}}(\boldsymbol{w})$, $w_d = g_{\boldsymbol{q}}(\boldsymbol{w})$ respectively where $\boldsymbol{w} = (w_1, \dots, w_{d-1})$. Also let $\mathcal{W}^{\|}$ be the tangent space of $\mathcal{P}_1^*$ at point $\boldsymbol{q} = \mathcal{P}_1^*(\theta)$ defined as the span of the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{q}} = \nabla_\theta \mathcal{P}_1^*(\theta)$, i.e.,

$$\mathcal{W}^{\|}(\boldsymbol{q}) = \left\{ \boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|-1} : \boldsymbol{v} = \sum_{i=1}^{\tilde{d}} \alpha_i \frac{\partial P_1^*(\theta)}{\partial \theta_i}, \alpha_i \in \mathbb{R} \right\}. \tag{40}$$

Note that if the Jacobian matrix is full rank then $\dim(\mathcal{W}^{\|}) = \tilde{d}$ which is equal to the dimension of the family of the distributions $\mathcal{P}_1$. However, in the case of rank deficient Jacobian matrix $\bar{d} \triangleq \dim(\mathcal{W}^{\|}) < \tilde{d}$, which means the projection of $\mathcal{P}_1$ into $\mathcal{B}(E_0)$ has less degrees of freedom than $\mathcal{P}_1$. Also, it can be shown that $\bar{d} \geq \tilde{d} - 1$. For every $\boldsymbol{q} \in \mathcal{P}_1^*$, we also define the orthogonal subspace $\mathcal{W}^{\perp}$ as

$$\begin{aligned} \mathcal{W}^{\perp}(\boldsymbol{q}) = \big\{ &\boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|-1} : \forall \boldsymbol{w} \in \mathcal{W}^{\|}(\boldsymbol{q}), \\ &\boldsymbol{v}^T \boldsymbol{w} = 0, \boldsymbol{v}^T \boldsymbol{n}(\boldsymbol{q}) = 0 \big\}. \end{aligned} \tag{41}$$

Hence every point in the tangent space of $\mathcal{B}(E_0)$ can be written as the direct sum of $\mathcal{W}^{\|}, \mathcal{W}^{\perp}$, i.e.,

$$\mathcal{W}^{\|}(\boldsymbol{q}) \oplus \mathcal{W}^{\perp}(\boldsymbol{q}) = \big\{\boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|-1} : \boldsymbol{v}^T \boldsymbol{n}(\boldsymbol{q}) = 0\big\}. \tag{42}$$

We can also decompose $\boldsymbol{w} = (\boldsymbol{w}^{\|}, \boldsymbol{w}^{\perp})$ such that $\boldsymbol{w}^{\|} \in \mathcal{W}^{\|}(\boldsymbol{q}), \boldsymbol{w}^{\perp} \in \mathcal{W}^{\perp}(\boldsymbol{q})$. Note that the decision region $\mathcal{Q}_1$ is union of hyperplanes tangent to the $\boldsymbol{q} \in \mathcal{P}_1^*$, and we have

$$\mathcal{Q}_1 \subseteq \bigcup_{\boldsymbol{q}\in\mathcal{P}_1^*} \mathcal{W}^{\perp}(\boldsymbol{q}) \oplus \mathcal{W}_d(\boldsymbol{q}), \tag{43}$$

where $\mathcal{W}_d(\boldsymbol{q}) = \text{span}\{\boldsymbol{n}(\boldsymbol{q})\}$. For every $\boldsymbol{q} \in \mathcal{P}_1^*$, let

$$\Gamma(\boldsymbol{q}) = \big\{(\boldsymbol{w}, w_d) : \boldsymbol{w} \in \Lambda(\boldsymbol{q}), w_d \geq g_{\boldsymbol{q}}(\boldsymbol{w})\big\}, \tag{44}$$

where

$$\Lambda(\boldsymbol{q}) = \Big\{\boldsymbol{w} : \frac{1}{2}\|\boldsymbol{n}(\boldsymbol{q})\|\boldsymbol{w}^T \boldsymbol{H}(\boldsymbol{q})\boldsymbol{w} \leq s\frac{\log n}{n}\Big\}, \tag{45}$$

and

$$\boldsymbol{H}(\boldsymbol{q}) = \nabla^2 g_{\boldsymbol{q}}(\boldsymbol{w})|_{\boldsymbol{w}=0} - \nabla^2 f_{\boldsymbol{q}}(\boldsymbol{w})|_{\boldsymbol{w}=0}, \tag{46}$$

and $s = \frac{d-\bar{d}-2}{2}$ chosen large enough such that the integral (38) over $\Gamma(\boldsymbol{q})$ dominates the integral over $\Gamma^c(\boldsymbol{q})$ for large n. Also note that in a neighbourhood of $\mathcal{W}^{\|}(\boldsymbol{q})$ we have $f(\boldsymbol{w}) = g(\boldsymbol{w})$, and for $\boldsymbol{w}^{\|} = \boldsymbol{0}$, $g_{\boldsymbol{q}}(\boldsymbol{w}) = 0$ for $\boldsymbol{w}^{\perp} \in \mathcal{W}^{\perp}(\boldsymbol{q})$. We limit the integral on the set $\Gamma = \bigcup_{\boldsymbol{q}\in\mathcal{P}_1^*} \Gamma(\boldsymbol{q})$ and further expand the integral (38) to get

$$\int_{\boldsymbol{q}\in\mathcal{Q}_1\cap\Gamma} e^{-nD(\boldsymbol{q}\|P_0)} d\boldsymbol{q} \leq \int_{\boldsymbol{q}\in\mathcal{P}_1^*} d\boldsymbol{w}^{\|} \int_{\boldsymbol{w}\in\mathcal{W}^{\perp}(\boldsymbol{q})\cap\Gamma(\boldsymbol{q})} d\boldsymbol{w}^{\perp} \int_{w_d=g_{\boldsymbol{q}}(\boldsymbol{w})}^{\infty} e^{-nD(\boldsymbol{q}\|P_0)} dw_d, \tag{47}$$

where we have dropped the dependence of $d\boldsymbol{w}^{\|}(\boldsymbol{q}), d\boldsymbol{w}^{\perp}(\boldsymbol{q})$ on $\boldsymbol{q}$ for ease of notation. Next by Taylor expanding $D\big(\boldsymbol{q} + (\boldsymbol{w}, w_d)\|P_0\big)$ for $\boldsymbol{q} \in \mathcal{P}_1^*$ in the d'th coordinate direction which is parallel to $\boldsymbol{n}(\boldsymbol{q})$ around $\boldsymbol{q} + \big(\boldsymbol{w}, f_{\boldsymbol{q}}(\boldsymbol{w})\big)$ we get

$$\begin{aligned} &D\big(\boldsymbol{q} + (\boldsymbol{w}, w_d)\|P_0\big) \geq D\big(\boldsymbol{q} + \big(\boldsymbol{w}, f_{\boldsymbol{q}}(\boldsymbol{w})\big)\|P_0\big) \\ &\quad + \nabla_{\boldsymbol{q}} D(\boldsymbol{q}\|P_0)^T\big|_{\boldsymbol{q}+(\boldsymbol{w},f_{\boldsymbol{q}}(\boldsymbol{w}))}(\boldsymbol{0}, w_d - f_{\boldsymbol{q}}(\boldsymbol{w})) && (48) \\ &= E_0 + \boldsymbol{n}\big(\boldsymbol{q} + (\boldsymbol{w}, f_{\boldsymbol{q}}(\boldsymbol{w}))\big)^T(\boldsymbol{0}, w_d - f_{\boldsymbol{q}}(\boldsymbol{w})) && (49) \\ &= E_0 + n_d(\boldsymbol{v})(w_d - f_{\boldsymbol{q}}(\boldsymbol{w})), && (50) \end{aligned}$$

where $n_d(\boldsymbol{v})$ is the projection of normal vector at point $\boldsymbol{q} + (\boldsymbol{w}, f_{\boldsymbol{q}}(\boldsymbol{w})$ onto the d'th coordinate, and in (48) we have used the convexity of relative entropy and hence convexity of relative entropy in all directions. Substituting the expansion in (47) we get

$$\begin{aligned} &\int_{\boldsymbol{w}\in\mathcal{Q}_1\cap\Gamma} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \leq e^{-nE_0} \int_{\boldsymbol{q}\in\mathcal{P}_1^*} d\boldsymbol{w}^{\|} \times \\ &\quad \int_{\boldsymbol{w}\in\mathcal{W}^{\perp}\cap\Gamma(\boldsymbol{q})} d\boldsymbol{w}^{\perp} \int_{w_d=g_{\boldsymbol{q}}(\boldsymbol{w})}^{\infty} e^{-nn_d(\boldsymbol{v})(w_d-f_{\boldsymbol{q}}(\boldsymbol{w}))} dw_d && (51) \\ &= \frac{1}{n} e^{-nE_0} \int_{\boldsymbol{q}\in\mathcal{P}_1^*} d\boldsymbol{w}^{\|} \times \\ &\quad \int_{\boldsymbol{w}^{\perp}\in\mathcal{W}^{\perp}\cap\Gamma(\boldsymbol{q})} \frac{1}{n_d(\boldsymbol{v})} e^{-nn_d(\boldsymbol{v})(g_{\boldsymbol{q}}(\boldsymbol{w})-f_{\boldsymbol{q}}(\boldsymbol{w}))} d\boldsymbol{w}^{\perp} && (52) \end{aligned}$$

Taylor expanding $n_d(\boldsymbol{v})$ and $g_{\boldsymbol{q}}(\boldsymbol{w}) - f_{\boldsymbol{q}}(\boldsymbol{w})$ over $\boldsymbol{w} = \boldsymbol{0}$ we get

$$g_{\boldsymbol{q}}(\boldsymbol{w}) - f_{\boldsymbol{q}}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^{\perp})^T \boldsymbol{H}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q})\boldsymbol{w}^{\perp} + o(\|\boldsymbol{w}\|^2) \tag{53}$$

$$n_d(\boldsymbol{q} + (\boldsymbol{w}, f_{\boldsymbol{q}}(\boldsymbol{w})) = \|\boldsymbol{n}(\boldsymbol{q})\| + o(\|\boldsymbol{w}\|), \tag{54}$$

where

$$\boldsymbol{H}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q}) = \nabla^2_{\boldsymbol{w}^{\perp}}\big(g_{\boldsymbol{q}}(\boldsymbol{w}) - f_{\boldsymbol{q}}(\boldsymbol{w})\big)\big|_{\boldsymbol{w}=0}, \tag{55}$$

is the $(d - \bar{d} - 1) \times (d - \bar{d} - 1)$ positive definite Hessian matrix of $g_{\boldsymbol{q}}(\boldsymbol{w}) - f_{\boldsymbol{q}}(\boldsymbol{w})$ in subspace $\mathcal{W}^{\perp}(\boldsymbol{q})$, and we used the fact that $f(\boldsymbol{w}) = g(\boldsymbol{w})$ in a neighbourhood $\boldsymbol{w} = (\boldsymbol{w}^{\|}, \boldsymbol{0})$ for $\|\boldsymbol{w}^{\|}\| \leq \epsilon$ for some positive ϵ. Therefore,

$$\begin{aligned} n_d(\boldsymbol{v})(g_{\boldsymbol{q}}(\boldsymbol{w}) - f_{\boldsymbol{q}}(\boldsymbol{w})) =& \frac{1}{2}\|\boldsymbol{n}(\boldsymbol{q})\|(\boldsymbol{w}^{\perp})^T \boldsymbol{H}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q})\boldsymbol{w}^{\perp} \\ &+ o(\|\boldsymbol{w}\|^2). \end{aligned} \tag{56}$$

Substituting in the integral (52), and by (45) we get

$$\begin{aligned} &\int_{\boldsymbol{w}\in\mathcal{Q}_1\cap\Gamma} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \leq \frac{1}{n} e^{-nE_0} \int_{\boldsymbol{q}\in\mathcal{P}_1^*} d\boldsymbol{w}^{\|} \times \\ &\int_{\boldsymbol{w}^{\perp}\in\mathcal{W}^{\perp}\cap\Gamma(\boldsymbol{q})} \frac{1}{\|\boldsymbol{n}(\boldsymbol{q})\|} e^{-n\frac{1}{2}\|\boldsymbol{n}(\boldsymbol{q})\|(\boldsymbol{w}^{\perp})^T \boldsymbol{H}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q})\boldsymbol{w}^{\perp}} d\boldsymbol{w}^{\perp} \\ &\qquad \times (1 + o(1)). \end{aligned} \tag{57}$$

By change of variable $\boldsymbol{u} = \sqrt{n}\boldsymbol{w}^{\perp}$, and hence $d\boldsymbol{u} = n^{\frac{d-\bar{d}-1}{2}} d\boldsymbol{w}^{\perp}$, we have

$$\begin{aligned} &\int_{\boldsymbol{w}\in\mathcal{Q}_1\cap\Gamma} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \leq n^{\frac{\bar{d}-d-1}{2}} e^{-nE_0} \int_{\boldsymbol{q}\in\mathcal{P}_1^*} \frac{1}{\|\boldsymbol{n}(\boldsymbol{q})\|} d\boldsymbol{w}^{\|} \\ &\times \int_{\mathbb{R}^{d-\bar{d}-1}} e^{-\frac{1}{2}\|\boldsymbol{n}(\boldsymbol{q})\|\boldsymbol{u}^T \boldsymbol{H}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q})\boldsymbol{u}} d\boldsymbol{u}(1 + o(1)), \end{aligned} \tag{58}$$

Where we have substituted the limits of the second integral to the whole space. The second integral in (58) is now a Gaussian integral for every $\boldsymbol{q}$, therefore

$$\begin{aligned} &\int_{\boldsymbol{w}\in\mathcal{Q}_1\cap\Gamma} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \leq n^{\frac{\bar{d}-d-1}{2}} e^{-nE_0} \times \\ &\int_{\boldsymbol{q}\in\mathcal{P}_1^*} \frac{\sqrt{2\pi\text{det}(\boldsymbol{H}^{-1}_{\boldsymbol{w}^{\perp}}(\boldsymbol{q}))}}{\|\boldsymbol{n}(\boldsymbol{q})\|^{\frac{3}{2}}} d\boldsymbol{w}^{\|}(\boldsymbol{q})(1 + o(1)) && (59) \\ &= n^{\frac{\bar{d}-d-1}{2}} e^{-nE_0}(c + o(1)), && (60) \end{aligned}$$

where in (60) we used that the integrand is independent of n and c is the integral of a function of the hessian matrix over the family $\mathcal{P}_1^*$. Approximating the remainder term, we have

$$\int_{\boldsymbol{w}\in\mathcal{Q}_1\cap\Gamma^c} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \leq \int_{\boldsymbol{w}\in\mathcal{B}(E_0+s\frac{\log n}{n})} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} \tag{61}$$

$$= n^{\frac{\bar{d}-d-1}{2}} e^{-nE_0} O(n^{-1}). \tag{62}$$

Finally, by (36), (38), and (60) we get (30).

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, July 2006.

[2] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, July 1974.

[3] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369–401, 04 1965.

[4] P. Boroumand and A. Guillén i Fàbregas, "Mismatched binary hypothesis testing: Error exponent sensitivity," *IEEE Trans. Inf. Theory*, pp. 1–1, 2022.

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

[6] I. Chakravarti and W. Hoeffding, *Asymptotic Theory of Statistical Tests and Estimation: In Honor of Wassily Hoeffding*, Academic Press. Academic Press, 1980.

[7] M. Iltis, "Sharp asymptotics of large deviations in $\mathbb{R}^d$," *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, 1995.

[8] G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Lancho, "Saddlepoint approximation of the error probability of binary hypothesis testing," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2306–2310.

[9] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, July 2014.

[10] I. Csiszár and P.C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[11] M. Walker, "A generalization of the maximum theorem," *International Economic Review*, vol. 20, no. 1, pp. 267–272, 1979.

[12] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1587–1603, 2011.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, March 2004.

[14] J.L. Jensen, *Saddlepoint Approximations*, Oxford science publications. Clarendon Press, 1995.

[15] R.W. Butler, *Saddlepoint Approximations with Applications*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2007.

[16] P. Ney, "Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$," *The Annals of Probability*, vol. 11, no. 1, pp. 158–167, 1983.