Error Exponents of Mismatched Likelihood Ratio Testing

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

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

***Abstract*—We study the problem of mismatched likelihood ratio test. We analyze the type-I and II error exponents when the actual distributions generating the observation are different from the distributions used in the test. We derive the worst-case error exponents when the actual distributions generating the data are within a relative entropy ball of the test distributions. In addition, we study the sensitivity of the test for small relative entropy balls.**

I. Introduction and Prelimenaries

Consider the binary hypothesis testing problem [1] where an observation $\boldsymbol{x} = (x_1, \dots, x_n)$ is generated from two possible distributions P_1^n and P_2^n defined on the probability simplex $\mathcal{P}(\mathcal{X}^n)$. We assume that P_1^n and P_2^n are product distributions, i.e., $P_1^n(\boldsymbol{x}) = \prod_{i=1}^n P_1(x_i)$, and similarly for P_2^n. For simplicity, we assume that both $P_1(x) > 0$ and $P_2(x) > 0$ for each $x \in \mathcal{X}$.

Let $\phi : \mathcal{X}^n \to \{1, 2\}$ be a hypothesis test that decides which distribution generated the observation $\boldsymbol{x}$. We consider deterministic tests ϕ that decide in favor of P_1^n if $\boldsymbol{x} \in \mathcal{A}_1$, where $\mathcal{A}_1 \subset \mathcal{X}^n$ is the decision region for the first hypothesis. We define $\mathcal{A}_2 = \mathcal{X}^n \setminus \mathcal{A}_1$ to be the decision region for the second hypothesis. The test performance is measured by the two possible pairwise error probabilities. The type-I and type-II error probabilities are defined as

$$\epsilon_1(\phi) = \sum_{\boldsymbol{x} \in \mathcal{A}_2} P_1^n(\boldsymbol{x}), \qquad \epsilon_2(\phi) = \sum_{\boldsymbol{x} \in \mathcal{A}_1} P_2^n(\boldsymbol{x}). \tag{1}$$

A hypothesis test is said to be optimal whenever it achieves the optimal error probability tradeoff given by

$$\alpha_\beta = \min_{\phi : \epsilon_2(\phi) \leq \beta} \epsilon_1(\phi). \tag{2}$$

The likelihood ratio test defined as

$$\phi_\gamma(\boldsymbol{x}) = \mathbb{1}\left\{ \frac{P_2^n(\boldsymbol{x})}{P_1^n(\boldsymbol{x})} \geq e^{n\gamma} \right\} + 1. \tag{3}$$

was shown in [2] to attain the optimal tradeoff (2). The type of a sequence $\boldsymbol{x} = (x_1, \dots, x_n)$ is $\hat{T}_{\boldsymbol{x}}(a) = \frac{N(a|\boldsymbol{x})}{n}$, where $N(a|\boldsymbol{x})$ is the number of occurrences of the symbol $a \in \mathcal{X}$ in the string. The likelihood ratio test can also be expressed as a function of the type of the observation $\hat{T}_{\boldsymbol{x}}$ as [3]

$$\phi_\gamma(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{ D(\hat{T}_{\boldsymbol{x}} \| P_1) - D(\hat{T}_{\boldsymbol{x}} \| P_2) \geq \gamma \big\} + 1. \tag{4}$$

where $D(P\|Q) = \sum_{\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the relative entropy between distributions P and Q.

In this paper, we are interested in the asymptotic exponential decay of the pairwise error probabilities. The optimal error exponent tradeoff (E_1, E_2) is defined as

$$\begin{aligned} E_2(E_1) \triangleq \sup \big\{ & E_2 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \\ & \epsilon_1(\phi) \leq e^{-nE_1} \quad \text{and} \quad \epsilon_2(\phi) \leq e^{-nE_2} \big\}. \end{aligned} \tag{5}$$

By using the Sanov's Theorem [3], [4], the optimal error exponent tradeoff (E_1, E_2), attained by the likelihood ratio test, can be shown to be [5], [6]

$$E_1(\phi_\gamma) = \min_{Q \in \mathcal{Q}_1(\gamma)} D(Q\|P_1), \tag{6}$$

$$E_2(\phi_\gamma) = \min_{Q \in \mathcal{Q}_2(\gamma)} D(Q\|P_2), \tag{7}$$

where

$$\mathcal{Q}_1(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_1) - D(Q\|P_2) \geq \gamma \big\}, \tag{8}$$

$$\mathcal{Q}_2(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_1) - D(Q\|P_2) \leq \gamma \big\}. \tag{9}$$

The minimizing distribution in (6), (7) is the tilted distribution

$$Q_\lambda(x) = \frac{P_1^{1-\lambda}(x) P_2^\lambda(x)}{\sum_{a \in \mathcal{X}} P_1^{1-\lambda}(a) P_2^\lambda(a)}, \quad 0 \leq \lambda \leq 1 \tag{10}$$

whenever γ satisfies $-D(P_1\|P_2) \leq \gamma \leq D(P_2\|P_1)$. In this case, λ is the solution of

$$D(Q_\lambda\|P_1) - D(Q_\lambda\|P_2) = \gamma. \tag{11}$$

Instead, if $\gamma < -D(P_1\|P_2)$, the optimal distribution in (6) is $Q_\lambda(x) = P_1(x)$ and $E_1(\phi_\gamma) = 0$, and if $\gamma > D(P_2\|P_1)$, the optimal distribution in (7) is $Q_\lambda(x) = P_2(x)$ and $E_2(\phi_\gamma) = 0$.

Equivalently, the dual expressions of (6) and (7) can be derived by substituting the minimizing distribution (10) into the Lagrangian yielding [4], [5]

$$E_1(\phi_\gamma) = \max_{\lambda \geq 0} \lambda\gamma - \log \Big(\sum_{x \in \mathcal{X}} P_1^{1-\lambda}(x) P_2^\lambda(x) \Big), \tag{12}$$

$$E_2(\phi_\gamma) = \max_{\lambda \geq 0} -\lambda\gamma - \log \Big(\sum_{x \in \mathcal{X}} P_1^\lambda(x) P_2^{1-\lambda}(x) \Big). \tag{13}$$

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 TEC2016-78434-C3-1-R.

The Stein regime is defined as the highest error exponent under one hypothesis when the error probability under the other hypothesis is at most some fixed $\epsilon \in (0, \frac{1}{2})$ [3]

$$E_2^{(\epsilon)} \triangleq \sup\big\{E_2 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \\ \epsilon_1(\phi) \le \epsilon \quad \text{and} \quad \epsilon_2(\phi) \le e^{-nE_2}\big\}. \tag{14}$$

The optimal $E_2^{(\epsilon)}$, given by [3]

$$E_2^{(\epsilon)} = D(P_1 \| P_2), \tag{15}$$

can be achieved by setting the threshold in (4) to be $\gamma = -D(P_1\|P_2) + \frac{C_2}{\sqrt{n}}$, where C_2 is a constant that depends on distributions P_1, P_2 and ϵ.

In this work, we revisit the above results in the case where the distributions used by the likelihood ratio test are not known precisely, and instead, fixed distributions $\hat{P}_1$ and $\hat{P}_2$ are used for testing. In particular, we find the error exponent tradeoff for fixed $\hat{P}_1$ and $\hat{P}_2$ and we study the worst-case tradeoff when the true distributions generating the observation are within a certain distance of the test distributions. The literature in robust hypothesis testing is vast (see e.g., [7]–[9] and references therein). Robust hypothesis testing consists of designing tests that are robust to the inaccuracy of the distributions generating the observation. Instead, we study the error exponent tradeoff performance of the likelihood ratio test for fixed test distributions. The proofs of our results can be found in [10].

II. Mismatched Likelihood Ratio Testing

Let $\hat{P}_1(x)$ and $\hat{P}_2(x)$ be the test distributions used in the likelihood ratio test with threshold $\hat{\gamma}$ given by

$$\hat{\phi}_{\hat{\gamma}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1) - D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2) \ge \hat{\gamma}\big\} + 1. \tag{16}$$

For simplicity, we assume that both $\hat{P}_1(x) > 0$ and $\hat{P}_2(x) > 0$ for each $x \in \mathcal{X}$. We are interested in the achievable error exponent of the mismatched likelihood ratio test, i.e.,

$$\hat{E}_2(\hat{E}_1) \triangleq \sup\big\{\hat{E}_2 \in \mathbb{R}_+ : \exists \hat{\gamma}, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \\ \epsilon_1(\hat{\phi}_{\hat{\gamma}}) \le e^{-n\hat{E}_1} \text{ and } \epsilon_2(\hat{\phi}_{\hat{\gamma}}) \le e^{-n\hat{E}_2}\big\}. \tag{17}$$

Theorem 1. *For fixed $\hat{P}_1, \hat{P}_2 \in \mathcal{P}(X)$ the optimal error exponent tradeoff in* (17) *is given by*

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = \min_{Q \in \hat{\mathcal{Q}}_1(\hat{\gamma})} D(Q\|P_1) \tag{18}$$

$$\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = \min_{Q \in \hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2) \tag{19}$$

where

$$\hat{\mathcal{Q}}_1(\hat{\gamma}) = \big\{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|\hat{P}_1) - D(Q\|\hat{P}_2) \ge \hat{\gamma}\big\}, \tag{20}$$

$$\hat{\mathcal{Q}}_2(\hat{\gamma}) = \big\{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|\hat{P}_1) - D(Q\|\hat{P}_2) \le \hat{\gamma}\big\}. \tag{21}$$

The minimizing distributions in (18) *and* (19) *are*

$$\hat{Q}_{\lambda_1}(x) = \frac{P_1(x)\hat{P}_1^{-\lambda_1}(x)\hat{P}_2^{\lambda_1}(x)}{\sum_{a\in\mathcal{X}} P_1(a)\hat{P}_1^{-\lambda_1}(a)\hat{P}_2^{\lambda_1}(a)}, \quad \lambda_1 \ge 0, \tag{22}$$

$$\hat{Q}_{\lambda_2}(x) = \frac{P_2(x)\hat{P}_2^{-\lambda_2}(x)\hat{P}_1^{\lambda_2}(x)}{\sum_{a\in\mathcal{X}} P_2(a)\hat{P}_2^{-\lambda_2}(a)\hat{P}_1^{\lambda_2}(a)}, \quad \lambda_2 \ge 0 \tag{23}$$

respectively, where λ_1 is chosen so that

$$D(\hat{Q}_{\lambda_1}\|\hat{P}_1) - D(\hat{Q}_{\lambda_1}\|\hat{P}_2) = \hat{\gamma}, \tag{24}$$

whenever $D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) \le \hat{\gamma}$, and otherwise, $\hat{Q}_{\lambda_1}(x) = P_1(x)$ and $\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = 0$. Similarly, $\lambda_2 \ge 0$ is chosen so that

$$D(\hat{Q}_{\lambda_2}\|\hat{P}_1) - D(\hat{Q}_{\lambda_2}\|\hat{P}_2) = \hat{\gamma}, \tag{25}$$

whenever $D(P_2\|\hat{P}_1) - D(P_2\|\hat{P}_2) \ge \hat{\gamma}$, and otherwise, $\hat{Q}_{\lambda_2}(x) = P_2(x)$ and $\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = 0$. Furthermore, the dual expressions for the type-I and type-II error exponents are

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = \max_{\lambda\ge 0} \lambda\hat{\gamma} - \log\Big(\sum_{x\in\mathcal{X}} P_1(x)\hat{P}_1^{-\lambda}(x)P_2^{\lambda}(x)\Big), \tag{26}$$

$$\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = \max_{\lambda\ge 0} -\lambda\hat{\gamma} - \log\Big(\sum_{x\in\mathcal{X}} P_1^{\lambda}(x)P_2(x)\hat{P}_2^{-\lambda}(x)\Big). \tag{27}$$

Remark 1*:* For mismatched likelihood ratio testing, the optimizing distributions $\hat{Q}_{\lambda_1}, \hat{Q}_{\lambda_2}$ can be different, since the decision regions only depend on the mismatched distributions. However, if $\hat{P}_1, \hat{P}_2$ are tilted with respect to P_1 and P_2, then both $\hat{Q}_{\lambda_1}, \hat{Q}_{\lambda_2}$ are also tilted respect to P_1 and P_2. This implies the result in [11], where for any set of mismatched distributions $\hat{P}_1, \hat{P}_2$ that are tilted with respect to generating distributions, the mismatched likelihood ratio test achieves the optimal error exponent tradeoff in (5).

Theorem 2. *In the Stein regime, the mismatched likelihood ratio test achieves*

$$\hat{E}_2^{(\epsilon)} = \min_{Q \in \hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2), \tag{28}$$

with threshold

$$\hat{\gamma} = D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) + \frac{\hat{C}_2}{\sqrt{n}}, \tag{29}$$

and $\hat{C}_2$ is a constant that depends on distributions $P_1, \hat{P}_1, \hat{P}_2$, and ϵ.

Remark 2*:* Note that since P_1 satisfies the constraint in (28) then $\hat{E}_2^{(\epsilon)} \le E_2^{(\epsilon)}$. In fact, if $\hat{P}_1, \hat{P}_2$ are tilted respect to P_1, P_2 then this inequality is met with equality. Moreover, it is easy to find a set of data and test distributions where $\hat{E}_2^{(\epsilon)} < E_2^{(\epsilon)}$.

III. Mismatched Likelihood Ratio Testing with Uncertainty

In this section, we analyze the worst-case error exponents tradeoff when the actual distributions P_1, P_2 are close to the mismatched test distributions $\hat{P}_1$ and $\hat{P}_2$. More specifically,

$$P_1 \in \mathcal{B}(\hat{P}_1, R_1), \quad P_2 \in \mathcal{B}(\hat{P}_2, R_2) \tag{30}$$

where the D-ball

$$\mathcal{B}(Q, R) = \big\{P \in \mathcal{P}(\mathcal{X}) : D(Q\|P) \le R\big\} \tag{31}$$

is a ball centered at distribution Q containing all distributions whose relative entropy is smaller or equal than radius R. This model was used in robust hypothesis testing in [12]. Figure 1 depicts the mismatched probability distributions and the mismatched likelihood ratio test as a hyperplane dividing the probability space into the two decision regions.

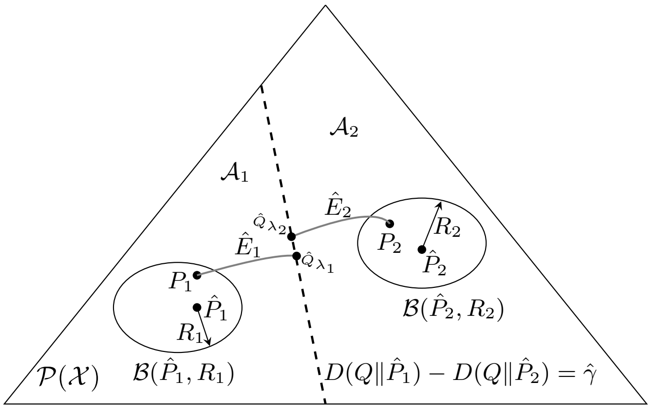

Fig. 1. Mismatched likelihood ratio test over distributions in D-balls.

We study the worst-case error-exponent performance of mismatched likelihood ratio testing when the distributions generating the observation fulfill (30). In particular, we are interested in the least favorable distributions P_1^L, P_2^L in $\mathcal{B}(\hat{P}_1, R_1), \mathcal{B}(\hat{P}_2, R_2)$, i.e., the distributions achieving the lowest error exponents $\hat{E}_1^L(R_1), \hat{E}_2^L(R_2)$.

Theorem 3. *For every $R_1, R_2 \geq 0$ let the least favorable exponents $\hat{E}_1^L(R_1), \hat{E}_2^L(R_2)$ defined as*

$$\hat{E}_1^L(R_1) = \min_{P_1 \in \mathcal{B}(\hat{P}_1, R_1)} \min_{Q \in \hat{\mathcal{Q}}_1(\hat{\gamma})} D(Q\|P_1), \tag{32}$$

$$\hat{E}_2^L(R_2) = \min_{P_2 \in \mathcal{B}(\hat{P}_2, R_2)} \min_{Q \in \hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2), \tag{33}$$

where $\hat{\mathcal{Q}}_1(\hat{\gamma}), \hat{\mathcal{Q}}_2(\hat{\gamma})$ are defined in (20), (21)*. Then, for any distribution pair $P_1 \in \mathcal{B}(\hat{P}_1, R_1), P_2 \in \mathcal{B}(\hat{P}_2, R_2)$, the corresponding error exponent pair $(\hat{E}_1, \hat{E}_2)$ satisfies*

$$\hat{E}_1^L(R_1) \leq \hat{E}_1(\hat{\phi}_{\hat{\gamma}}), \quad \hat{E}_2^L(R_2) \leq \hat{E}_2(\hat{\phi}_{\hat{\gamma}}). \tag{34}$$

Furthermore, the optimization problem in (32) *is convex with optimizing distributions*

$$Q_{\lambda_1}^L(x) = \frac{P_1^L(x)\hat{P}_1^{-\lambda_1}(x)\hat{P}_2^{\lambda_1}(x)}{\sum_{a \in \mathcal{X}} P_1^L(a)\hat{P}_1^{-\lambda_1}(a)\hat{P}_2^{\lambda_1}(a)}, \tag{35}$$

$$P_1^L(x) = \beta_1 Q_{\lambda_1}^L(x) + (1-\beta_1)\hat{P}_1(x), \tag{36}$$

where $\lambda_1 \geq 0, 0 \leq \beta_1 \leq 1$ are chosen such that

$$D(Q_{\lambda_1}^L\|\hat{P}_1) - D(Q_{\lambda_1}^L\|\hat{P}_2) = \hat{\gamma}, \tag{37}$$

$$D(\hat{P}_1\|P_1^L) = R_1, \tag{38}$$

when

$$\max_{P_1 \in \mathcal{B}(\hat{P}_1, R_1)} D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) \leq \hat{\gamma}. \tag{39}$$

Otherwise, we can find a least favorable distribution $P_1^L \in \mathcal{B}(\hat{P}_1, R_1)$ such that $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ for this distribution is $\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = 0$. Similarly, the optimization (33) *is convex with optimizing distributions*

$$Q_{\lambda_2}^L(x) = \frac{P_2^L(x)\hat{P}_2^{-\lambda_2}(x)\hat{P}_1^{\lambda_2}(x)}{\sum_{a \in \mathcal{X}} P_2^L(a)\hat{P}_2^{-\lambda_2}(a)\hat{P}_1^{\lambda_2}(a)}, \tag{40}$$

$$P_2^L(x) = \beta_2 Q_{\lambda_2}^L(x) + (1-\beta_2)\hat{P}_2(x), \tag{41}$$

where $\lambda_2 \geq 0, 0 \leq \beta_2 \leq 1$ are chosen such that

$$D(Q_{\lambda_2}^L\|\hat{P}_2) - DQ_{\lambda_2}^L\|\hat{P}_1) = \hat{\gamma}, \tag{42}$$

$$D(\hat{P}_2\|P_2^L) = R_2, \tag{43}$$

whenever,

$$\min_{P_2 \in \mathcal{B}(\hat{P}_2, R_2)} D(P_2\|\hat{P}_1) - D(P_2\|\hat{P}_2) \geq \hat{\gamma}. \tag{44}$$

Otherwise, we can find a distribution $P_2^L \in \mathcal{B}(\hat{P}_2, R_2)$ such that $\hat{E}_2(\hat{\phi}_{\hat{\gamma}})$ for this distribution is $\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = 0$.

The worst-case achievable error exponents of mismatched likelihood ratio testing for data distributions in a D-ball are essentially the minimum relative entropy between two sets of probability distributions. Specifically, the minimum relative entropy $\mathcal{B}(\hat{P}_1, R_1)$ and $\hat{\mathcal{Q}}_2(\hat{\gamma})$ gives $\hat{E}_1^L(R_1)$, and similarly for $\hat{E}_2^L(R_2)$.

IV. Mismatched Likelihood Ratio Testing Sensitivity

In this section, we study how the worst-case error exponents $(\hat{E}_1^L, \hat{E}_2^L)$ behave when the D-ball radii R_1, R_2 are small. In particular, we derive a Taylor series expansion of the worst-case error exponent. This approximation can also be interpreted as the worst-case sensitivity of the test, i.e., how does the test perform when actual distributions are very close to the mismatched distributions.

Theorem 4. *For every $R_i \geq 0$, $\hat{P}_i \in \mathcal{P}(\mathcal{X})$ for $i = 1, 2$, and*

$$-D(\hat{P}_1\|\hat{P}_2) \leq \hat{\gamma} \leq D(\hat{P}_2\|\hat{P}_1), \tag{45}$$

let

$$\tilde{E}_i^L(R_i) = E_i(\hat{\phi}_{\hat{\gamma}}) - S_i(\hat{P}_1, \hat{P}_2, \hat{\gamma})\sqrt{R_i}, \tag{46}$$

where

$$S_i^2(\hat{P}_1, \hat{P}_2, \hat{\gamma}) = 2Var_{\hat{P}_i}\left(\frac{\hat{Q}_\lambda(X)}{\hat{P}_i(X)}\right) \tag{47}$$

and $\hat{Q}_\lambda(X)$ is the minimizing distribution in (10) *for test $\hat{\phi}_{\hat{\gamma}}$. Then, approximation* (46) *is accurate for small R_i, i.e., for $R_i \to 0$*

$$\tilde{E}_i^L(R_i) \to \hat{E}_i^L(R_i). \tag{48}$$

Lemma 5. *For every $\hat{P}_1, \hat{P}_2 \in \mathcal{P}(\mathcal{X})$, and $\hat{\gamma}$ satisfying* (45)

$$\frac{\partial}{\partial\hat{\gamma}} S_1(\hat{P}_1, \hat{P}_2, \hat{\gamma}) \geq 0, \quad \frac{\partial}{\partial\hat{\gamma}} S_2(\hat{P}_1, \hat{P}_2, \hat{\gamma}) \leq 0. \tag{49}$$

This lemma shows that $S_1(\hat{P}_1, \hat{P}_2, \hat{\gamma})$ is a non-decreasing function of $\hat{\gamma}$, i.e., as $\hat{\gamma}$ increases from $-D(\hat{P}_1\|\hat{P}_2)$ to

$D(\hat{P}_2\|\hat{P}_1)$, the worst-case exponent $\hat{E}_1^L(R_1)$ becomes more sensitive to mismatch with likelihood ratio testing. Conversely, $S_2(\hat{P}_1,\hat{P}_2,\hat{\gamma})$ is a non-increasing function of $\hat{\gamma}$, i.e., as $\hat{\gamma}$ increases from $-D(\hat{P}_1\|\hat{P}_2)$ to $D(\hat{P}_2\|\hat{P}_1)$, the worst-case exponent $\hat{E}_2^L(R_2)$ becomes less sensitive (more robust) to mismatch with likelihood ratio testing. Moreover, when $\lambda=\frac{1}{2}$, we have

$$\hat{Q}_{\frac{1}{2}}(x)=\frac{\sqrt{\hat{P}_1(x)\hat{P}_2(x)}}{\sum_{a\in\mathcal{X}}\sqrt{\hat{P}_1(a)\hat{P}_2(a)}}, \tag{50}$$

and then $S_1(\hat{P}_1,\hat{P}_2,\hat{\gamma})=S_2(\hat{P}_1,\hat{P}_2,\hat{\gamma})$. In addition, $\hat{Q}_{\frac{1}{2}}$ minimizes $E_1(\hat{\phi}_{\hat{\gamma}})+E_2(\hat{\phi}_{\hat{\gamma}})$ yielding [13]

$$\begin{aligned} E_1(\hat{\phi}_{\hat{\gamma}})+E_2(\hat{\phi}_{\hat{\gamma}}) &= \min_{Q\in\mathcal{P}(\mathcal{X})} D(Q\|\hat{P}_1)+D(Q\|\hat{P}_2) \qquad (51)\\ &= 2B(\hat{P}_1,\hat{P}_2) \qquad (52)\end{aligned}$$

where $B(\hat{P}_1,\hat{P}_2)$ is the Bhattacharyya distance between the mismatched distributions $\hat{P}_1$ and $\hat{P}_2$. This suggests that having equal sensitivity (or robustness) for both hypotheses minimizes the sum of the exponents.

Example 1. When $\gamma=0$ the likelihood ratio test becomes the maximum-likelihood test, which is known to achieve the lowest average probability of error in the Bayes setting for equal priors. For fixed priors π_1,π_2, the error probability in the Bayes setting is $\bar{\epsilon}=\pi_1\epsilon_1+\pi_2\epsilon_2$, resulting in the following error exponent [3]

$$\bar{E}=\lim_{n\to\infty}\frac{1}{n}\log\bar{\epsilon}=\min\{E_1,E_2\}. \tag{53}$$

Consider $\hat{P}_1=\text{Bern}(0.1)$, $\hat{P}_2=\text{Bern}(0.8)$. Also, assume $R_1=R_2=R$. Figure 2 shows the worst-case error exponent in the Bayes setting given by $\min\{\hat{E}_1^L,\hat{E}_2^L\}$ by solving (32) and (33) as well as $\min\{\tilde{E}_1^L,\tilde{E}_2^L\}$ using the approximation in (46). We can see that the approximation is good for small R. Moreover, it can be seen that error exponents are very sensitive to mismatch for small R, i.e., the slope of the worst-case exponent goes to infinity as R approaches to zero.

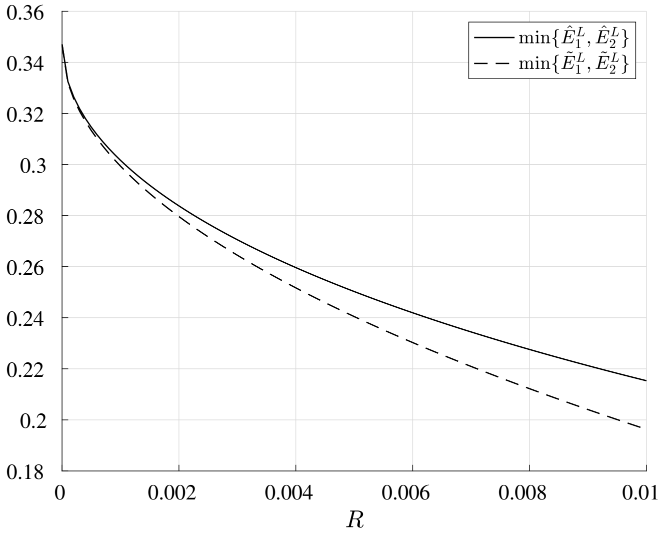

Fig. 2. Worst-case achievable Bayes error exponent.

Appendix

A. Proof of Theorem 1

We show the result for $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ and similar steps are valid for $\hat{E}_2(\hat{\phi}_{\hat{\gamma}})$. The type-I probability of error can be written as

$$\hat{\epsilon}_1(\hat{\phi}_{\hat{\gamma}})=\sum_{\substack{\boldsymbol{x}\in\mathcal{X}^n\\ D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1)-D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2)\geq\hat{\gamma}}} P_1^n(\boldsymbol{x}). \tag{54}$$

Applying Sanov's Theorem to (54) to get (18) is immediate. The optimization problem in (18) consists of the minimization of a convex function over linear constraints. Therefore, the KKT conditions are also sufficient [14]. Writing the Lagrangian, we have

$$\begin{aligned} L(Q,\lambda,\nu)=&D(Q\|P_1)+\lambda\big(D(Q\|\hat{P}_2)-D(Q\|\hat{P}_1)+\hat{\gamma}\big)\\ &+\nu\Big(\sum_{x\in\mathcal{X}}Q(x)-1\Big). \qquad (55)\end{aligned}$$

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

$$1+\log\frac{Q(x)}{P_1(x)}+\lambda\log\frac{\hat{P}_1(x)}{\hat{P}_2(x)}+\nu=0. \tag{56}$$

Solving equations (56) for every $x\in\mathcal{X}$ we obtain (22). Moreover, from the complementary slackness condition if [14]

$$D(P_1\|\hat{P}_1)-D(P_1\|\hat{P}_2)\leq\hat{\gamma}, \tag{57}$$

then (24) should hold. Otherwise, if (57) does not hold then λ in (56) should be zero and hence $\hat{Q}_{\lambda_1}=P_1$, $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})=0$. Finally, substituting the minimizing distribution $\hat{Q}_{\lambda_1}$ (22) into (55) we get the dual expression

$$g(\lambda)=\lambda\hat{\gamma}-\log\Big(\sum_{x\in\mathcal{X}}P_1\hat{P}_1^{-\lambda}(x)P_2^{\lambda}(x)\Big). \tag{58}$$

Since the optimization problem in (18) is convex, then the duality gap is zero [14], and this proves the (26).

B. Proof of Theorem 2

First, notice that $\hat{E}_2(\hat{\phi}_{\hat{\gamma}})$ is a non-increasing function of $\hat{\gamma}$ since for every $\hat{\gamma}_1\leq\hat{\gamma}_2$ we have

$$\hat{\mathcal{Q}}_2(\hat{\gamma}_1)\subset\hat{\mathcal{Q}}_2(\hat{\gamma}_2), \tag{59}$$

hence

$$\hat{E}_2(\hat{\phi}_{\hat{\gamma}_2})\leq\hat{E}_2(\hat{\phi}_{\hat{\gamma}_1}). \tag{60}$$

Therefore, in the Stein's regime we are looking for the smallest threshold such that $\limsup_{n\to\infty}\hat{\epsilon}_1(\hat{\phi}_{\hat{\gamma}})\leq\epsilon$. Let

$$\hat{\gamma}=D(P_1\|\hat{P}_1)-D(P_1\|\hat{P}_2)-\sqrt{\frac{V(P_1,\hat{P}_1,\hat{P}_2)}{n}}\Phi^{-1}(\epsilon), \tag{61}$$

where

$$V(P_1, \hat{P}_1, \hat{P}_2) = \mathrm{Var}_{P_1}\bigg(\log \frac{\hat{P}_1}{\hat{P}_2} \bigg)$$
$$= \sum_{x \in \mathcal{X}} P_1(x) \bigg(\log \frac{\hat{P}_1}{\hat{P}_2} \bigg)^2 - \big(D(P_1\|\hat{P}_2) - D(P_1\|\hat{P}_1) \big)^2, \tag{62}$$

and $\Phi^{-1}(\epsilon)$ is the inverse cumulative distribution function of a zero-mean unit-variance Guassian random variable. For such $\hat{\gamma}$, the type-I error probability of the mismatched likelihood ratio test is

$$\hat{\epsilon}_1(\hat{\phi}_{\hat{\gamma}}) = \mathbb{P}_1 \Bigg[\frac{1}{n} \sum_{i=1}^{n} \log \frac{\hat{P}_1(X_i)}{\hat{P}_2(X_i)} \leq D(P_1\|\hat{P}_2) - D(P_1\|\hat{P}_1) + \sqrt{\frac{V(P_1, \hat{P}_1, \hat{P}_2)}{n}} \Phi^{-1}(\epsilon) \Bigg]. \tag{63}$$

Observe that $D(P_1\|\hat{P}_2) - D(P_1\|\hat{P}_1) = \mathbb{E}_{P_1}\Big[\log \frac{\hat{P}_1(X)}{\hat{P}_2(X)} \Big]$. Let $\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\imath}(x_i)$, where $\hat{\imath}(x_i) = \log \frac{\hat{P}_1(x_i)}{\hat{P}_2(x_i)}$. Letting Z be a zero-mean unit-variance Guassian random variable, then, by the central limit theorem we have

$$\limsup_{n\to\infty} \hat{\epsilon}_1(\hat{\phi}_{\hat{\gamma}})$$
$$= \limsup_{n\to\infty} \mathbb{P}_1 \Bigg[\frac{\sqrt{n}\big(\hat{S}_n - \mathbb{E}_{P_1}[\hat{\imath}(X)]\big)}{\sqrt{V(P_1, \hat{P}_1, \hat{P}_2)}} \leq \Phi^{-1}(\epsilon) \Bigg] \tag{64}$$
$$= \mathbb{P}\big[Z \leq \Phi^{-1}(\epsilon) \big] \tag{65}$$
$$= \epsilon. \tag{66}$$

Therefore, asymptotically, the type-I error probability of mismatched likelihood ratio test with $\hat{\gamma}$ in (61) is equal to ϵ.

Next, we need to show that for any threshold $\hat{\gamma}$ and $\varepsilon > 0$ such that

$$\limsup_{n\to\infty} \hat{\gamma} + \varepsilon \leq D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2), \tag{67}$$

the type-I probability of error tends to 1 as the number of observation approaches infinity, which implies that $D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2)$ is the lowest possible threshold that meets the constraint $\limsup_{n\to\infty} \hat{\epsilon}_1(\hat{\phi}_{\hat{\gamma}}) \leq \epsilon$. The corresponding $\hat{E}_2(\hat{\phi}_{\hat{\gamma}})$ is this highest type-II exponent that meets the constraint. In order to show this, define the following sets

$$\mathcal{E}_\delta = \Big\{ \boldsymbol{x} \in \mathcal{X}^n : \|\hat{T}_{\boldsymbol{x}}(x) - P_1(x)\|_\infty < \delta \Big\}, \tag{68}$$
$$\mathcal{D} = \big\{ \boldsymbol{x} \in \mathcal{X}^n : \big| D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1) - D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2) - D(P_1\|\hat{P}_1) + D(P_1\|\hat{P}_2) \big| < \varepsilon \big\}, \tag{69}$$
$$\bar{\mathcal{D}} = \big\{ \boldsymbol{x} \in \mathcal{X} : D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1) - D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2) - D(P_1\|\hat{P}_1) + D(P_1\|\hat{P}_2) \geq -\varepsilon \big\}. \tag{70}$$

where $\|.\|_\infty$ is the norm infinity. From the continouity of $D(.\|\hat{P})$ we have that for any $\varepsilon > 0$ such that

$$\big| D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1) - D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2) - D(P_1\|\hat{P}_1) + D(P_1\|\hat{P}_2) \big| < \varepsilon. \tag{71}$$

there exists $\delta > 0$ such that for all $\hat{T}_{\boldsymbol{x}}$ satisfying

$$\|\hat{T}_{\boldsymbol{x}}(x) - P_1(x)\|_\infty < \delta \tag{72}$$

(71) holds. Therefore, when (67) holds

$$\liminf_{n\to\infty} \epsilon_1(\hat{\phi}_{\hat{\gamma}}) \geq \liminf_{n\to\infty} \sum_{x \in \bar{\mathcal{D}}} P_1^n(\boldsymbol{x}) \tag{73}$$
$$\geq \liminf_{n\to\infty} \sum_{x \in \mathcal{D}} P_1^n(\boldsymbol{x}). \tag{74}$$

Now from the continuity argument, there exists a δ such that

$$\sum_{x \in \mathcal{D}} P_1^n(\boldsymbol{x}) \geq \sum_{x \in \mathcal{E}_\delta} P_1^n(\boldsymbol{x}). \tag{75}$$

Set $\delta_n = \sqrt{\frac{\log n}{n}}$. Thus, for sufficiently large n, $\delta_n \leq \delta$, Therefore, we have

$$\liminf_{n\to\infty} \epsilon_1(\hat{\phi}_{\hat{\gamma}}) \geq \liminf_{n\to\infty} \sum_{\boldsymbol{x} \in \mathcal{E}_{\delta_n}} P_1^n(\boldsymbol{x}) \tag{76}$$
$$\geq \lim_{n\to\infty} 1 - \frac{2|\mathcal{X}|}{n} \tag{77}$$
$$= 1. \tag{78}$$

where the last step is by Hoeffding's inequality [15] and union bound. Therefore, for any $\hat{\gamma} < D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2)$ type-I error goes to unity which concludes the theorem.

C. Proof of Theorem 3

We show the result under the first hypothesis and similar steps are valid under the second hypothesis. For every P_1 the achievable type-I is error exponent $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ does not depend on P_2 therefore, (32) is a lower bound to $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$. Moreover, since the relative entropy is jointly convex, then (32) is a convex optimization problem and the KKT conditions are also sufficient. Writing the Lagrangian we have

$$L(Q, P_1, \lambda_1, \lambda_1', \nu_1, \nu_1') = D(Q\|P_1) + \lambda_1\big(D(Q\|\hat{P}_2) - D(Q\|\hat{P}_1) + \hat{\gamma}\big) + \lambda_1'\big(D(\hat{P}_1\|P_1) - R_1\big) + \nu_1\Big(\sum_{x\in\mathcal{X}} Q(x) - 1\Big) + \nu_1'\Big(\sum_{x\in\mathcal{X}} P_1(x) - 1\Big). \tag{79}$$

Differentiating with respect to $Q(x)$ and $P_1(x)$ and setting the derivatives to zero we have

$$1 + \log \frac{Q(x)}{P_1(x)} + \lambda_1 \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} + \nu_1 = 0, \tag{80}$$
$$-\frac{Q(x)}{P_1(x)} - \lambda_1' \frac{\hat{P}_1(x)}{P_1(x)} + \nu_1' = 0, \tag{81}$$

respectively. Solving equations (80), (81) for every $x \in \mathcal{X}$ and letting $\beta_1 = \frac{1}{1+\lambda_1'}$ we obtain (35) and (36). Moreover, from the complementary slackness condition [14] if for all P_1 in $\mathcal{B}(\hat{P}_1, R_1)$ the condition $D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) \leq \hat{\gamma}$ stated in Theorem 1 holds, then (37) and (38) should hold. Otherwise, if there exists a P_1^L in $\mathcal{B}(\hat{P}_1, R_1)$ such that $D(P_1^L\|\hat{P}_1) - D(P_1^L\|\hat{P}_2) \leq \hat{\gamma}$, then for this distribution $\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = 0$. Therefore, if conditon (39) holds for all P_1 in the D-ball $\hat{E}_1^L(R_1) > 0$, otherewise $\hat{E}_1^L(R_1) = 0$.

D. Proof of Theorem 4

We show the result under the first hypothesis, and similar steps are valid for the second hypothesis. First, we consider the first minimization in (32) over Q, i.e.,

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = \min_{Q \in \hat{\mathcal{Q}}_1(\hat{\gamma})} D(Q\|P_1) \tag{82}$$

By the envelope theorem [16] we have

$$\frac{\partial \hat{E}_1(\hat{\phi}_{\hat{\gamma}})}{\partial P_1(x)} = -\frac{\hat{Q}_\lambda(x)}{P_1(x)}. \tag{83}$$

Define the vectors

$$\nabla \hat{E}_1 = \left(-\frac{\hat{Q}_\lambda(x_1)}{\hat{P}_1(x_1)}, \dots, -\frac{\hat{Q}_\lambda(x_{|\mathcal{X}|})}{\hat{P}_1(x_{|\mathcal{X}|})} \right)^T \tag{84}$$

$$\boldsymbol{\theta}_{P_1} = \big(P_1(x_1) - \hat{P}_1(x_1), \dots, P_1(x_{|\mathcal{X}|}) - \hat{P}_1(x_{|\mathcal{X}|})\big)^T. \tag{85}$$

By applying a Taylor expansion to $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ around $P_1 = \hat{P}_1$ we obtain

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = E_1(\hat{\phi}_{\hat{\gamma}}) + \boldsymbol{\theta}_{P_1}^T \nabla \hat{E}_1 + o(\|\boldsymbol{\theta}_{P_1}\|_\infty). \tag{86}$$

By substituting the expansion (86) for the first minimization in (32) we obtain

$$\hat{E}_1^L(R_1) = \min_{P_1 \in \mathcal{B}(\hat{P}_1, R_1)} E_1(\hat{\phi}_{\hat{\gamma}}) + \boldsymbol{\theta}_{P_1}^T \nabla \hat{E}_1 + o(\|\boldsymbol{\theta}_{P_1}\|_\infty). \tag{87}$$

Now, we further approximate the outer minimization constraint in (32). By y approximating $D(\hat{P}_1\|P_1)$ we get [17]

$$D(\hat{P}_1\|P_1) = \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} + o(\|\boldsymbol{\theta}_{P_1}\|_\infty^2), \tag{88}$$

where

$$\boldsymbol{J}(\hat{P}_1) = \operatorname{diag}\left(\frac{1}{\hat{P}_1(x_1)}, \dots, \frac{1}{\hat{P}_1(x_{|\mathcal{X}|})}\right) \tag{89}$$

is the Fisher information matrix []. Therefore, (87) can be approximated as

$$\hat{E}_1^L(R_1) \approx \tilde{E}_1^L(R_1) \triangleq \min_{\substack{\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} \le R_1 \\ \boldsymbol{1}^T \boldsymbol{\theta}_{P_1} = 0}} E_1(\hat{\phi}_{\hat{\gamma}}) + \boldsymbol{\theta}_{P_1}^T \nabla \hat{E}_1. \tag{90}$$

The optimization problem in (90) is convex and hence the KKT conditions are sufficient. The corresponding Lagrangian is given by

$$L(\boldsymbol{\theta}_{P_1}, \lambda, \nu) = E_1(\hat{\phi}_{\hat{\gamma}}) + \boldsymbol{\theta}_{P_1}^T \nabla \hat{E}_1 + \lambda\Big(\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} - R_1\Big) + \nu(\boldsymbol{1}^T\boldsymbol{\theta}_{P_1}). \tag{91}$$

Differentiating with respect to $\boldsymbol{\theta}_{P_1}$ and setting to zero, we have

$$\nabla \hat{E}_1 + \lambda \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} + \nu \boldsymbol{1} = 0. \tag{92}$$

Therefore,

$$\boldsymbol{\theta}_{P_1} = -\frac{1}{\lambda}\boldsymbol{J}^{-1}(\hat{P}_1)\big(\nabla \hat{E}_1 + \nu \boldsymbol{1}\big). \tag{93}$$

Note that if $\lambda = 0$ then from (92) $\nabla \hat{E}_1 = -\nu \boldsymbol{1}$ which cannot be true for thresholds satisfying (45) since $\hat{Q}_\lambda \neq \hat{P}_1$. Therefore, from the complementary slackness condition [14] the inequality constraint (90) should be satisfied with equality. By solving $\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} = R_1$ and $\boldsymbol{1}^T\boldsymbol{\theta}_{P_1} = 0$ and substituting λ, ν in (93), we obtain

$$\boldsymbol{\theta}_{P_1} = -\frac{\boldsymbol{\psi}}{\sqrt{\boldsymbol{\psi}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\psi}}}\sqrt{2R_1}, \tag{94}$$

where

$$\boldsymbol{\psi} = \boldsymbol{J}^{-1}(\hat{P}_1)\Bigg(\nabla \hat{E}_1 - \boldsymbol{1}^T\boldsymbol{J}^{-1}(\hat{P}_1)\nabla \hat{E}_1 \boldsymbol{1}\Bigg). \tag{95}$$

Substituding (94) into (90) yields (46).

Finally, we prove the continuity statement. Intuitively, we need to prove that as R_1 goes to zero, then all the approximations above become exact. First, note that by assumption, $\hat{P}_1(x) > 0$ for each $x \in \mathcal{X}$. Therefore, for any finite R_1, we have $P_1(x) > 0$ for every $P_1 \in \mathcal{B}(\hat{P}_1, R_1)$. Hence, for $P_1 \in \mathcal{B}(\hat{P}_1, R_1)$, the relative entropy $D(Q\|P_1)$ is continuous in both Q, P_1. Moreover, the constraints in (82) are continuous with respect to Q and also trivially with respect to P_1, since the constraints do not depend on P_1. Hence, the optimization in (82) is minimizing a continuous function over a compact set with continuous constraints. Hence, by the maximum theorem [18], $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ is a continuous function of P_1 for all $P_1 \in \mathcal{B}(\hat{P}_1, R_1)$ with finite radius R_1. Moreover, by (94), $\|\boldsymbol{\theta}_{P_1}\|_\infty \to 0$ as $R_1 \to 0$. Hence as $R \to 0$, we have the following:

- $E_1(\hat{\phi}_{\hat{\gamma}}) + \boldsymbol{\theta}_{P_1}^T \nabla_{P_1} \hat{E}_1 \to \hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ by (86) and continuity of $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ for $P_1 \in \mathcal{B}(\hat{P}_1, R_1), R_1 < \infty$.
- $\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}(\hat{P}_1)\boldsymbol{\theta}_{P_1} \to D(\hat{P}_1\|P_1)$ by (88) and the continuity of $D(\hat{P}_1\|P_1)$ for small R_1.

Finally, by the continuity of the objective function in (90) with respect to the constraints, we conclude (48).

E. Proof of Lemma 5

We show the result under the first hypothesis and similar steps are valid under the second hypothesis. To prove the Theorem we need the following lemma.

Lemma 6. *Consider the following optimization problem*

$$E(\gamma) = \min_{\mathbb{E}_Q[X] \ge \gamma} D(Q\|P). \tag{96}$$

Then $E(\gamma)$ is convex in γ.

Proof: Let

$$Q_1^* = \operatorname*{arg\,min}_{\mathbb{E}_Q[X] \ge \gamma_1} D(Q\|P) \quad Q_2^* = \operatorname*{arg\,min}_{\mathbb{E}_Q[X] \ge \gamma_2} D(Q\|P). \tag{97}$$

From the convexity of the relative entropy, for any $\alpha \in (0, 1)$,

$$D\big(\alpha Q_1^* + (1-\alpha)Q_2^* \| P\big) \le \alpha D(Q_1^*\|P) + (1-\alpha)D(Q_2^*\|P) \tag{98}$$

$$= \alpha \min_{\mathbb{E}_Q[X]\ge\gamma_1} D(Q\|P) + (1-\alpha)\min_{\mathbb{E}_Q[X]\ge\gamma_2} D(Q\|P). \tag{99}$$

Furthermore, since Q_1^*, Q_2^* satisfy their correspending optimization constraints, then $\mathbb{E}_{Q_1^*}[X]\ge\gamma_1, \mathbb{E}_{Q_2^*}[X]\ge\gamma_2$, hence

$$\mathbb{E}_{\alpha Q_1^*+(1-\alpha)Q_2^*}[X] \ge \alpha\gamma_1+(1-\alpha)\gamma_2. \tag{100}$$

Therefore, $\alpha Q_1^* + (1-\alpha)Q_2^*$ satisfies the optimization constraint when $\gamma=\alpha\gamma_1+(1-\alpha)\gamma_2$, then

$$\min_{\mathbb{E}_Q[X]\le\alpha\gamma_1+(1-\alpha)\gamma_2} D(Q\|P) \le D(\alpha Q_1^*+(1-\alpha)Q_2^*\|P) \tag{101}$$

$$\le \alpha \min_{\mathbb{E}_Q[X]\ge\gamma_1} D(Q\|P) + (1-\alpha)\min_{\mathbb{E}_Q[X]\ge\gamma_2} D(Q\|P). \tag{102}$$

Hence $E(\gamma)$ is convex in γ.

■

From above lemma we can show that λ is a non-decreasing function of $\hat\gamma$. From the envelope theorem [16]

$$\frac{\partial \hat E_1(\hat\phi_{\hat\gamma})}{\partial\hat\gamma} = \lambda^*, \tag{103}$$

where λ^* is the optimizing λ in (10) for the test $\hat\phi_{\hat\gamma}$. Therefore

$$\frac{\partial\lambda^*}{\partial\hat\gamma} = \frac{\partial^2 \hat E_1(\hat\phi_{\hat\gamma})}{\partial\hat\gamma^2} \ge 0, \tag{104}$$

where the inequality is from convexity of $\hat E_1(\hat\phi_{\hat\gamma})$ respect to $\hat\gamma$. Therefore, we only need to consider the behavior of variance as λ changes. Taking the derivative of variance respect to λ, we have

$$\frac{\partial}{\partial\lambda}\mathrm{Var}_{\hat P_1}\Big(\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\Big) = \sum_{x\in\mathcal{X}} \frac{2\hat Q_\lambda(x)}{\hat P_1(x)}\frac{\partial \hat Q_\lambda(x)}{\partial\lambda} \tag{105}$$

$$\begin{aligned} &= \sum_{x\in\mathcal{X}} \frac{2\hat Q_\lambda(x)}{\hat P_1(x)}\Bigg(\hat Q_\lambda(x)\log\frac{\hat P_2(x)}{\hat P_1(x)} - \\ &\qquad \hat Q_\lambda(x)\sum_{x'\in\mathcal{X}}\hat Q_\lambda(x')\log\frac{\hat P_2(x')}{\hat P_1(x')}\Bigg) \end{aligned} \tag{106}$$

$$\begin{aligned} &= 2\mathbb{E}_{\hat Q_\lambda}\left[\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\log\frac{\hat P_2(X)}{\hat P_1(X)}\right] \\ &\quad - 2\mathbb{E}_{\hat Q_\lambda}\left[\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\right]\mathbb{E}_{\hat Q_\lambda}\left[\log\frac{\hat P_2(X)}{\hat P_1(X)}\right]. \end{aligned} \tag{107}$$

Substituting $\hat Q_\lambda(X)$ as a function of λ we get

$$\begin{aligned} &\frac{\sum_{a\in\mathcal{X}}\hat P_1^{1-\lambda}(a)\hat P_2^{\lambda}(a)}{2}\frac{\partial}{\partial\lambda}\mathrm{Var}_{\hat P_1}\left(\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\right) \\ &= \mathbb{E}_{\hat Q_\lambda}\left[\left(\frac{\hat P_2(X)}{\hat P_1(X)}\right)^{\lambda}\log\frac{\hat P_2(X)}{\hat P_1(X)}\right] \\ &\quad - \mathbb{E}_{\hat Q_\lambda}\left[\left(\frac{\hat P_2(X)}{\hat P_1(X)}\right)^{\lambda}\right]\mathbb{E}_{\hat Q_\lambda}\left[\log\frac{\hat P_2(X)}{\hat P_1(X)}\right]. \end{aligned} \tag{108}$$

Let $r(X) = \left(\frac{\hat P_2(X)}{\hat P_1(X)}\right)^{\lambda}$, then

$$\begin{aligned} &\mathbb{E}_{\hat Q_\lambda}\left[\left(\frac{\hat P_2(X)}{\hat P_1(X)}\right)^{\lambda}\log\frac{\hat P_2(X)}{\hat P_1(X)}\right] \\ &\qquad - \mathbb{E}_{\hat Q_\lambda}\left[\left(\frac{\hat P_2(X)}{\hat P_1(X)}\right)^{\lambda}\right]\mathbb{E}_{\hat Q_\lambda}\left[\log\frac{\hat P_2(X)}{\hat P_1(X)}\right] \\ &= \frac{1}{\lambda}\mathbb{E}_{\hat Q_\lambda}[r(X)\log r(X)] - \frac{1}{\lambda}\mathbb{E}_{\hat Q_\lambda}[r(X)]\mathbb{E}_{\hat Q_\lambda}[\log r(X)]. \end{aligned} \tag{109}$$

Note that $\hat Q_\lambda(x), r(x)$ are positive for all $x\in\mathcal{X}$. Therefore, using the log-sum inequality [3] for the first term and Jensen inequality [3] for the second term in (109), we obtain

$$\begin{aligned} &\frac{\lambda\sum_{a\in\mathcal{X}}\hat P_1^{1-\lambda}(a)\hat P_2^{\lambda}(a)}{2}\frac{\partial}{\partial\lambda}\mathrm{Var}_{\hat P_1}\Big(\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\Big) \\ &\ge \mathbb{E}_{\hat Q_\lambda}[r(X)]\log\mathbb{E}_{\hat Q_\lambda}[r(X)] - \mathbb{E}_{\hat Q_\lambda}[r(X)]\mathbb{E}_{\hat Q_\lambda}[\log r(X)] \end{aligned} \tag{110}$$

$$\ge \mathbb{E}_{\hat Q_\lambda}[r(X)]\log\mathbb{E}_{\hat Q_\lambda}[r(X)] - \mathbb{E}_{\hat Q_\lambda}[r(X)]\log\mathbb{E}_{\hat Q_\lambda}[r(X)] \tag{111}$$

$$= 0. \tag{112}$$

Also, the above inequalities are met with equality when both log-sum and Jensen's inequalities are met with equality, which happens when $\lambda=0$. Therefore, for $\lambda>0$, $\mathrm{Var}_{\hat P_1}\left(\frac{\hat Q_\lambda(X)}{\hat P_1(X)}\right)$ is an increasing function of λ for $\lambda>0$ and consequently

$$\frac{\partial}{\partial\hat\gamma}S_1(\hat P_1,\hat P_2,\hat\gamma) \ge 0. \tag{113}$$

References

[1] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, Springer Texts in Statistics. Springer, New York, third edition, 2005.

[2] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

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

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 95, 01 2010.

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

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

[7] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, 12 1965.

[8] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, March 1985.

[9] H. V. Poor, *An introduction to signal detection and estimation*, Springer, 2013.

[10] P. Boroumand and A. Guillén i Fàbregas, "Error exponents of mismatched likelihood ratio testing," `https://tinyurl.com/y2vgvj4u`, 2019.

[11] 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, Mar. 2011.

[12] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 413–421, Jan 2009.

[13] 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.

[14] S. Boyd and L.Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[15] W. Hoeffding, "Probability inequalities for sums of bounded random variables," 1962.

[16] P Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

[17] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE Int. Zürich Seminar on Commun.*, March 2008, pp. 14–17.

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