Information and Inference: A Journal of the IMA (2024) **13**, iaae010
https://doi.org/10.1093/imaiai/iaae010

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Universal Neyman–Pearson classification with a partially known hypothesis

Parham Boroumand*
Department of Engineering, University of Cambridge, Trumpington Street, CB2 1PZ, Cambridge, UK
*Corresponding author: Email: pb702@cam.ac.uk

and

Albert Guillén i Fàbregas
Department of Engineering, University of Cambridge, Trumpington Street, CB2 1PZ, Cambridge, UK and Department of Information and Communications Technologies, Universitat Pompeu Fabra, C/Roc Boronat 138, 08018 Barcelona, Spain

[Received on 7 June 2022; revised on 26 December 2023; accepted on 9 April 2024]

We propose a universal classifier for binary Neyman–Pearson classification where the null distribution is known, while only a training sequence is available for the alternative distribution. The proposed classifier interpolates between Hoeffding's classifier and the likelihood ratio test and attains the same error probability prefactor as the likelihood ratio test, i.e. the same prefactor as if both distributions were known. In addition, such as Hoeffding's universal hypothesis test, the proposed classifier is shown to attain the optimal error exponent tradeoff attained by the likelihood ratio test whenever the ratio of training to observation samples exceeds a certain value. We propose a lower bound and an upper bound to the optimal training to observation ratio. In addition, we propose a sequential classifier that attains the optimal error exponent tradeoff.

Keywords: classification; composite hypothesis testing; sequential hypothesis testing; Neyman–Pearson hypothesis testing; likelihood ratio test.

1. Introduction

The problem of deciding between multiple statistical descriptions of a given observation is one of the main problems in statistics and finds applications in a number of fields including engineering, signal processing, medicine and finance among others [24]. Depending on whether or not the possible probability distributions of the observation are known, the problem is termed hypothesis testing or classification. In the case where there are only two possible distributions of the observation, one typically refers to these decision problems as binary. When priors on the distributions are available, the problem is cast as Bayesian and the average probability of error determines the quality of the detection. In the absence of priors, the design of tests and classifiers minimizes one pairwise error probability while keeping the other upper bounded by a constant. This setting, proposed by Neyman and Pearson [30], has been adopted as the *de-facto* test design setting since it allows for both pairwise error probabilities to be bounded, unlike average risk minimization.

Upon observing a vector of n observations, the hypothesis test that minimizes the above pairwise error probability tradeoff when the two possible distributions are known is given by the likelihood ratio test [30]. Instead, when decisions need to be made each time a new observation arrives, the sequential probability ratio test (SPRT) provides the best exponential decay of the error probability (or error exponent) [26, 39]. In this case, the decision time is a random variable which is constrained in mean. When one of the hypothesis distributions belongs to some class of distributions and only the class is

© The Author(s) 2024. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

known, Hoeffding proposed the generalized likelihood ratio test that attains the optimal error exponent [18]. When the distributions are not known, several classifiers, such as logistic regression, support vector machines and naive Bayes, have been proposed in the literature [9, 10, 15, 17]. However, none of these classifiers provides a guarantee on the type-I error probability resulting in the possibility of a large type-I errors. In the Neyman–Pearson setting, [35] gives uniform bounds on type-I error probability for the plugin likelihood ratio test when both distributions are unknown, and only training sequences from both distributions are available. Gutman [16] proposed universal classifiers that guarantee a certain type-I error exponent under any probability distribution pair while achieving the lowest type-II error probability. Asymptotic refinements, including second, third and fourth order, of the error exponent analysis have been developed in the data transmission context in [28, 31]. Zhou *et al.* [42] proved the second order optimality of Gutman's classifier. Zeitouni *et. al* [41] extended Hoeffding's test to the case where both hypotheses are known to belong to a parametric family of distributions. The Kolmogorov–Smirnov test [21] can also be used in the case when the null hypothesis is known and has continuous alphabet.

This paper considers the setting when the null hypothesis generating distribution is known, while only a training sequence is available for the alternative hypothesis. This can be the case in the unbalanced training sample size when a large number of training samples is available for the null hypothesis so that with high accuracy, the distribution can be estimated. In contrast, the alternative distribution cannot be accurately estimated due to small number of training samples. The scenario when the number of test and training samples is fixed can also be viewed as composite hypothesis testing with known null hypothesis and an additional training sequence. Hoeffding's test can achieve the optimal error exponent when the null hypothesis distribution is given. However, the prefactor attained by Hoeffding's test is only optimal for distributions defined over binary alphabets.

In this paper, we propose a classifier that achieves the optimal prefactor (up to a constant factor) using the additional training samples while attaining the optimal error exponent tradeoff whenever the training to observation sample ratio exceeds a certain value. The proposed classifier interpolates between the plugin likelihood ratio and Hoeffding's tests. We also study the case when the test and training samples are generated sequentially. There is no classifier similar to Hoeffding's in this scenario that can achieve the error exponent gain by taking test samples sequentially when only the null hypothesis is known. However, the proposed classifier can improve the achievable error exponent by exploiting the additional training samples from the unknown distribution.

This paper is structured as follows. Section 2 introduces notation and reviews the preliminaries about the likelihood ratio test and Hoeffding's generalized likelihood ratio test. Section 3 proposes the new classifier for a fixed sample size setting and shows the finite length improvements over Hoeffding's test. Section 4 discusses the classification problem in the Stein regime and shows the optimality of Gutman's test in this regime. Section 5 proposes a sequential classifier achieving the highest error exponent tradeoff in the proposed setting. Proofs of the main results can be found in the Appendices.

2. Prelimenaries

Consider the following binary classification 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 the probability simplex $\mathscr{P}(\mathscr{X})$ with alphabet size $|\mathscr{X}| < \infty$. We assume that the distribution P_0 is known, while only a sequence of training samples $\boldsymbol{z} = (z_1, \dots, z_k) \sim P_1^k$ generated in an i.i.d. fashion from P_1 is available; training and test sequences are sampled independently from each other. We also assume that both $P_0(x) > 0, P_1(x) > 0$ and $\frac{P_0(x)}{P_1(x)} \leq c$ for each $x \in \mathscr{X}$ for some positive c. Also we let k, the length of the training sequence, be such that $k = \alpha n$ for some positive α.

The type of an n-length sequence $\boldsymbol{y}$ is defined as $\hat{T}_{\boldsymbol{y}}(a) = \frac{N(a|\mathbf{y})}{n}$, where $N(a|\boldsymbol{y})$ is the number of occurrences of symbol $a \in \mathscr{X}$ in sequence $\boldsymbol{y}$. The types of the observation and training sequences $\boldsymbol{x}, \boldsymbol{z}$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

are denoted by $\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}$, respectively. The set of all sequences of length n with type P, denoted by $\mathscr{T}_P^n$, is called the type class. The set of types formed with length n sequences on the simplex $\mathscr{P}(\mathscr{X})$ is denoted as $\mathscr{P}_n(\mathscr{X})$.

Let $\phi(\boldsymbol{z}, \boldsymbol{x}) : \mathscr{X}^k \times \mathscr{X}^n \to \{0, 1\}$ be a classifier that decides the distribution that generated the observation $\boldsymbol{x}$ upon processing the training sequence $\boldsymbol{z}$. We consider deterministic classifiers ϕ that decide in favor of P_0 if $\boldsymbol{x} \in \mathscr{A}_0(P_0, \boldsymbol{z})$, where $\mathscr{A}_0(P_0, \boldsymbol{z}) \subset \mathscr{X}^n$ is the decision region for the first hypothesis and is a function of P_0 and the training samples $\boldsymbol{z}$. We define $\mathscr{A}_1(P_0, \boldsymbol{z}) = \mathscr{X}^n \backslash \mathscr{A}_0$ to be the decision region for the second hypothesis. If we assume no prior knowledge on either distribution, the two possible pairwise error probabilities determine the performance of the classifier. Specifically, the type-I and type-II error probabilities are defined as

$$\varepsilon_0(\phi) = \sum_{\boldsymbol{z} \in \mathscr{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathscr{A}_1(P_0, \boldsymbol{z})} P_0(\boldsymbol{x}), \tag{2.1}$$

$$\varepsilon_1(\phi) = \sum_{\boldsymbol{z} \in \mathscr{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathscr{A}_0(P_0, \boldsymbol{z})} P_1(\boldsymbol{x}). \tag{2.2}$$

In the case where both distributions are known, the training sequence is not needed and the classifier becomes a hypothesis test. In this case, the classifier is said to be optimal whenever it achieves the optimal error probability tradeoff given by

$$\min_{\phi : \varepsilon_0(\phi) \leq \xi} \varepsilon_1(\phi), \tag{2.3}$$

where $\xi \in [0, 1]$. It is well known that likelihood ratio test defined as

$$\phi^{\mathrm{lrt}}(\boldsymbol{x}) = \mathbb{1}\left\{ \frac{P_1^n(\boldsymbol{x})}{P_0^n(\boldsymbol{x})} \geq e^{n\gamma} \right\} \tag{2.4}$$

attains the optimal tradeoff (2.3) for every γ. This is the well-known Neyman–Pearson lemma [30]. The likelihood ratio test can also be expressed as a function of the type of the observation $\hat{T}_{\boldsymbol{x}}$ as [11, 13]

$$\phi^{\mathrm{lrt}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{ D(\hat{T}_{\boldsymbol{x}} \| P_0) - D(\hat{T}_{\boldsymbol{x}} \| P_1) \geq \gamma \big\}, \tag{2.5}$$

where $D(P\|Q) = \sum_{x \in \mathscr{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the relative entropy between distributions P and Q. The optimal error exponent tradeoff (E_0, E_1) is defined as

$$E_1^*(E_0) \triangleq \sup \big\{ E_1 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0, \varepsilon_0(\phi) \leq e^{-nE_0} \quad \text{and} \quad \varepsilon_1(\phi) \leq e^{-nE_1} \big\}. \tag{2.6}$$

In the case where both E_0, E_1 are strictly positive, we refer to the error exponent tradeoff in (2.6) as the Chernoff regime. By using Sanov's Theorem [11, 14], the optimal error exponent tradeoff (E_1, E_0) attained by the likelihood ratio is given by

$$E_0(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathscr{Q}_0(\gamma)} D(Q\|P_0), \tag{2.7}$$

$$E_1(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathscr{Q}_1(\gamma)} D(Q\|P_1), \tag{2.8}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where

$$\mathcal{Q}_0(\gamma) = \{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \geq \gamma\}, \tag{2.9}$$

$$\mathcal{Q}_1(\gamma) = \{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \leq \gamma\}. \tag{2.10}$$

By varying the threshold γ in the range $-D(P_0\|P_1) \leq \gamma \leq D(P_1\|P_0)$, eqs (2.7) and (2.8) fully characterize the error exponent tradeoff in (2.6). Furtheremore, the minimizing distribution in (2.7), (2.8) is the tilted distribution

$$Q_{\lambda^*}(x) = \frac{P_0^{\lambda^*}(x)P_1^{1-\lambda^*}(x)}{\sum_{a\in\mathcal{X}} P_0^{1-\lambda^*}(a)P_1^{\lambda^*}(a)}, \quad 0 \leq \lambda^* \leq 1, \tag{2.11}$$

where λ^* is the solution of $D(Q_{\lambda^*}\|P_0) - D(Q_{\lambda^*}\|P_1) = \gamma$.

The classification problem described above with known P_0 and a training sequence from P_1, can also be viewed as a composite binary hypothesis problem where additional training sequence samples are given for the second hypotheses. In the case of a composite hypothesis testing problem where P_0 is given, and the other hypothesis is unrestricted to $\mathcal{P}(\mathcal{X})$, Hoeffding proposed in [18] a generalized likelihood-ratio test given by

$$\phi^{\text{glrt}}(\boldsymbol{x}) = \mathbb{1}\{D(\hat{T}_{\boldsymbol{x}}\|P_0) > E_0\}, \tag{2.12}$$

which attains the optimal error exponent tradeoff in (2.6). By Sanov's theorem, the error exponent of Hoeffding's test is given by

$$E_0(\phi^{\text{glrt}}) = E_0, \tag{2.13}$$

$$E_1(\phi^{\text{glrt}}) = \min_{\substack{Q\in\mathcal{P}(\mathcal{X}),\\ D(Q\|P_0)\leq E_0}} D(Q\|P_1), \tag{2.14}$$

where the minimizing distributions in (2.14) is given by

$$Q_{\mu^*}(x) = \frac{P_0^{\frac{\mu^*}{1+\mu^*}}(x)P_1^{\frac{1}{1+\mu^*}}(x)}{\sum_{a\in\mathcal{X}} P_0^{\frac{\mu^*}{1+\mu^*}}(a)P_1^{\frac{1}{1+\mu^*}}(a)}, \quad \mu^* \geq 0, \tag{2.15}$$

and μ^* is the solution to $D(Q_{\mu^*}\|P_0) = E_0$. By varying the threshold E_0 in the range $0 \leq E_0 \leq D(P_1\|P_0)$, eqs (2.13) and (2.14) fully characterize the error exponent tradeoff in (2.6). Using a large deviations refinement [20, 37], the type-I error probability of the likelihood ratio test is

$$\varepsilon_0(\phi^{\text{lrt}}) = \frac{1}{\sqrt{n}} e^{-nE_0}\big(c + o(1)\big), \tag{2.16}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

while Hoeffding's test type-I error probability is given by [20, 25]

$$\varepsilon_0(\phi^{\mathrm{glrt}}) = n^{\frac{|\mathscr{X}|-3}{2}} e^{-nE_0}\big(c' + o(1)\big), \tag{2.17}$$

where c, c' are constants that only depend on P_0, P_1 and the corresponding test thresholds. Since the likelihood ratio and Hoeffding's tests attain the optimal error exponent tradeoff (2.6), for any fixed E_0, then $E_1(\phi^{\mathrm{glrt}}) = E_1(\phi^{\mathrm{lrt}})$. As a result, when the number of observations is large, Hoeffding's test, although attaining the optimal error exponent tradeoff, suffers in exponential prefactor when compared with the likelihood ratio's $\frac{1}{\sqrt{n}}$ for observation alphabets such that $|\mathscr{X}| > 2$. For $|\mathscr{X}| = 2$, the decision regions for the likelihood ratio and Hoeffding's tests coincide, and thus, (2.17) is the same as (2.16).

3. Fixed sample sized universal classifier

In this section, we propose a classifier that interpolates between the likelihood ratio and Hoeffding's tests that attains a prefactor that is independent of the alphabet size and is equal to $\frac{1}{\sqrt{n}}$. In addition, we show that if the ratio of training samples to the number of test samples α exceeds a certain threshold, the proposed test also achieves the optimal error exponent tradeoff.

Hoeffding's test can favor the second hypothesis for test sequences with types close to P_0 while far from P_1. Suppose we have a training sequence type $\hat{T}_{z}$, we can relax the Hoeffding's test from a ball centered at P_0 to a hyperplane tangent to the Hoeffding's test ball, directed toward the type of the training sequence. As we will see, this is precisely what enables the improvement in the prefactor of the type-I probability of error. We propose the following classifier:

$$\phi_\beta(\hat{T}_{\boldsymbol{x}}, \hat{T}_{z}) = \mathbb{1}\big\{\beta D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_{z}) - D(\hat{T}_{\boldsymbol{x}}\|P_0) \ \leq \gamma(E_0, \hat{T}'_{z})\big\}, \tag{3.1}$$

where $0 \leq \beta \leq 1$, the threshold $\gamma(E_0, Q_1)$ is given by

$$\gamma(E_0, Q_1) = \beta \min_{\substack{Q\in\mathscr{P}(\mathscr{X}),\\ D(Q\|P_0)\leq E_0}} D(Q\|Q_1) - E_0, \tag{3.2}$$

the perturbed training type $\hat{T}'_{z}(a)$ is

$$\hat{T}'_{z}(a) = \big(1-\delta_n\big)\hat{T}_{z}(a) + \frac{\delta_n}{|\mathscr{X}|}, \tag{3.3}$$

where δ_n can be chosen as any function of the order $o(n^{-1})$. We add this small perturbation of the training type to avoid the cases where some of the alphabet symbols have not been observed in the training sequence. We define the decision regions of the proposed classifier by

$$\mathscr{A}_0(\hat{T}_{z}, \beta) = \{Q : Q \in \mathscr{P}(\mathscr{X}), \phi_\beta(Q, \hat{T}_{z}) = 0\}, \tag{3.4}$$

$$\mathscr{A}_1(\hat{T}_{z}, \beta) = \{Q : Q \in \mathscr{P}(\mathscr{X}), \phi_\beta(Q, \hat{T}_{z}) = 1\}. \tag{3.5}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

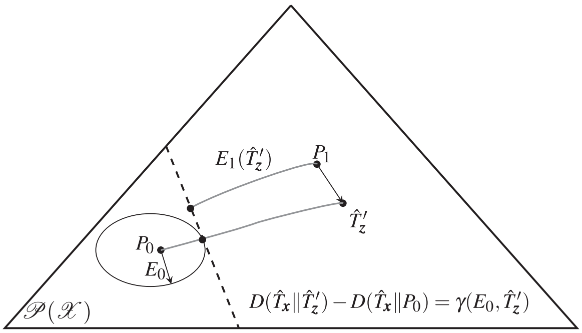
$E_1(\hat{T}_z')$
P_1
$\hat{T}_z'$
P_0
E_0
$\mathscr{P}(\mathscr{X})$
$D(\hat{T}_x\|\hat{T}_z') - D(\hat{T}_x\|P_0) = \gamma(E_0,\hat{T}_z')$

FIG. 1. Proposed classifier with known distribution P_0 and a training sequence with type $\hat{T}_z$.

Since parameter β controls how much the training weights in the decision, we have that when $\beta = 0$, we recover Hoeffding's test, while for $\beta = 1$, the test is reminiscent of a likelihood ratio test where instead of P_1, we have the perturbed training type $\hat{T}_z'(a)$. Intuitively, as long as we have enough training samples, the training type $\hat{T}_z'(a)$ will be close to P_1 and we will attain the optimal error exponent tradeoff. This is indeed what will be shown next.

Figure 1 illustrates the proposed classifier for a realization of the training sequence when $\beta = 1$. The proposed classifier becomes the plug-in likelihood ratio test, where the test's threshold has been adjusted by the training samples 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. The type-I error exponent will be equal to E_0 for any realization of $\hat{T}_z$. However, the type-II error exponent for a given training sample $E_1(\hat{T}_z')$ is the projection of P_1 into the separating hyperplane determined by the test, which is the function of the training sequence and can vary accordingly.

Next, we find a refined expression for the type-I error probability and show that the error probability prefactor is $O\big(\frac{1}{\sqrt{n}}\big)$, i.e. of the same order of the prefactor achieved by the likelihood ratio test.

THEOREM 3.1 For $P_0, P_1, 0 < \beta \leq 1$ and fixed E_0, the classifier ϕ_β defined in (3.1) attains a type-I error probability such that

$$\varepsilon_0(\phi_\beta) = \frac{1}{\sqrt{n}} e^{-nE_0}(c + o(1)). \tag{3.6}$$

In addition, for every $P_0, P_1, E_0, \beta \in (0, 1]$, there exists a finite training to sample size ratio α_β^* such that for any $\alpha > \alpha_\beta^*$, we have that

$$\varepsilon_1(\phi_\beta) = \frac{1}{\sqrt{n}} e^{-nE_1^*(E_0)}(c' + o(1)), \tag{3.7}$$

where c, c' are positive constants that only depend on the data distributions and E_0.

Proof. The proof can be found in Appendix B. □

We have shown that the classifier proposed in (3.1) not only achieves the optimal error exponent tradeoff for $\alpha > \alpha_\beta^*$ but also achieves the same prefactor of the type-I error probability of the likelihood ratio test. This represents a significant improvement with respect to the Hoeffding's universal test for

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

observation alphabets $|\mathscr{X}| > 2$. In addition, we show that the proposed classifier achieves the same prefactor as the type-II error probability of the likelihood ratio test, establishing the optimality of the proposed classifier up to a constant.

Example 3.2 We present a numerical example to illustrate the performance of the proposed classifier in (3.1). Consider two trinary distributions $P_0 = [0.3, 0.3, 0.4], P_1 = [0.35, 0.35, 0.3]$ and set $\alpha = 2$ and $E_0 = 0.005$. 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 classifier in (3.1) as well as those of the likelihood ratio test in (2.5) and Hoeffding's classifier over $5 \cdot 10^7$ independent experiments. As it can be seen in Fig. 2, the type-I error exponent of the likelihood ratio test, and the proposed classifier are very close to each other and outperform the Hoeffding's test. In addition, we observe that the type-II error exponent of the proposed classifier is slightly worse than that of the likelihood ratio and Hoeffding's tests for small n; as n increases, the proposed classifier achieves the optimal exponent. In addition, in order to clearly observe the effect of the prefactor, in Fig. 3, we plot $\log \frac{\varepsilon_i(\phi)}{\frac{1}{\sqrt{n}}e^{-nE_i}}$ for $i \in \{0, 1\}$ for the same classifiers. We first notice that in the case of $\log \frac{\varepsilon_1(\phi)}{\frac{1}{\sqrt{n}}e^{-nE_1}}$, the three classifiers achieve the optimal $\frac{1}{\sqrt{n}}$ prefactor, though with different constants. We also observe that $\log \frac{\varepsilon_0(\phi)}{\frac{1}{\sqrt{n}}e^{-nE_0}}$ converges to a constant for the likelihood ratio test and the new classifier, as predicted by our analysis. Instead, as (2.17) suggests, Hoeffding's classifier fails to attain the optimal $\frac{1}{\sqrt{n}}$ prefactor for ε_0.

Theorem 3.1 shows that the classifier in (3.1) can achieve the optimal error exponent tradeoff if the training to sample ratio α is large enough. As it is evident by the proof of Theorem 3.1, the proposed classifier cannot achieve the optimal error exponent tradeoff for α close to zero. Therefore, it is desirable to find the smallest α^*_β such that the above theorem holds. The next results introduce lower and upper bounds to α^*_β. As it is apparent from the proof, the proposed bounds are specific to the proposed classifier.

Before stating the next result, we need some additional definitions. Define the matrix

$$\boldsymbol{H} = \beta\eta^*_\beta\sqrt{\boldsymbol{J}}\Big[\boldsymbol{Q} + \eta^*_1\boldsymbol{V} + (1 - \eta^*_1)\boldsymbol{W} - \boldsymbol{T}\Big]\sqrt{\boldsymbol{J}}, \tag{3.8}$$

where

$$\boldsymbol{T} = \mathrm{diag}(Q_{\mu^*}), \quad \boldsymbol{J} = \mathrm{diag}\left(\frac{1}{P_1}\right), \tag{3.9}$$

where Q_{μ^*} is defined in (2.15) when $Q_1 = P_1$, $\boldsymbol{W} = \boldsymbol{w}\boldsymbol{w}^T$, $\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^T$, $\boldsymbol{Q} = \boldsymbol{q}\boldsymbol{q}^T$, where for $i = 1, \ldots, |\mathscr{X}|$, the entries of $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{q}$ are, respectively, defined as

$$w_i = \frac{1}{\sqrt{\mathrm{Var}_{Q_{\mu^*}}\left(\log \frac{Q_{\mu^*}}{P_0}\right)}} Q_{\mu^*}(i)\log\left(\frac{Q_{\mu^*}(i)}{P_0(i)}\right) - E_0 \tag{3.10}$$

$$v_i = \frac{1}{\sqrt{\mathrm{Var}Q_{\eta^*_\beta}(\varOmega)}} Q_{\eta^*_\beta}(i)\varOmega(i) \tag{3.11}$$

$$q_i = Q_{\mu^*}(i), \tag{3.12}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

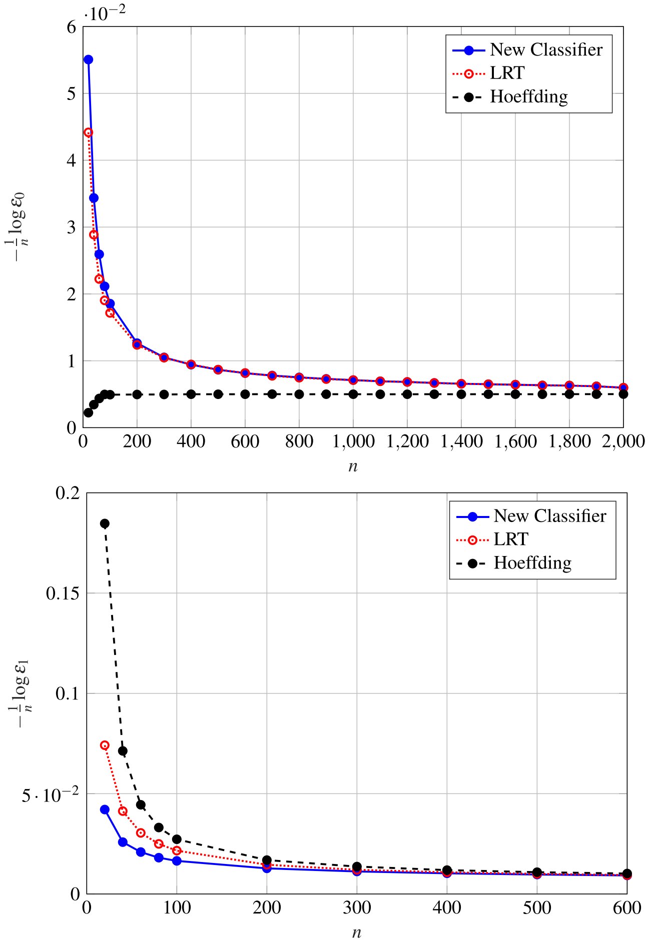

FIG. 2. Type-I and type-II error exponents for the likelihood ratio test, Hoeffding's test and the proposed classifier.

where $Q_{\eta^*_\beta}$, is the projection of P_1 onto $\mathscr{A}_0(P_1,\beta)$, which equals to Q_{μ^*}, i.e. the tilted distribution

$$Q_{\eta^*_\beta}(x) = \frac{P_1^{\frac{1-\eta^*_\beta\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(x)P_0^{\frac{\eta^*_\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(x)}{\sum_{a\in\mathscr{X}}P_1^{\frac{1-\eta^*_\beta\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(a)P_0^{\frac{\eta^*_\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(a)}, \tag{3.13}$$

$$\varOmega(i) = \beta\log\frac{P_1(i)}{P_0(i)} + (1-\beta)\log\frac{Q_{\eta^*_\beta}(i)}{P_0(i)}, \quad i\in\mathscr{X}, \tag{3.14}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

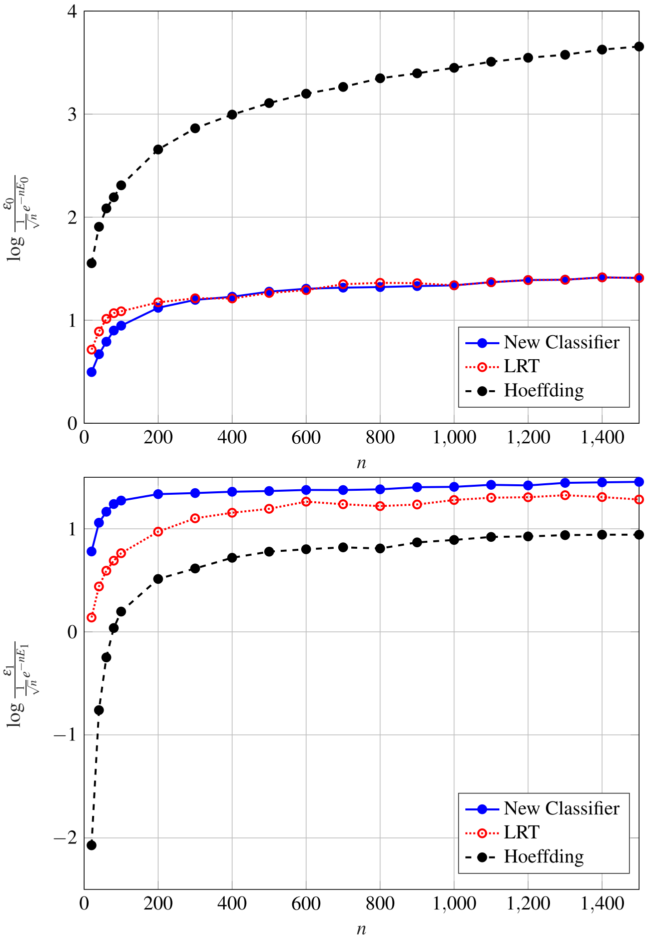

Fig. 3. log $\frac{\varepsilon_i(\phi)}{\frac{1}{\sqrt{n}}e^{-nE_i}}$ for $i \in \{0, 1\}$ and for the likelihood ratio test, Hoeffding's test and the proposed classifier.

and η^*_β is the optimal Lagrange multiplier in the problem below for $0 < \beta \leq 1$

$$L(Q, Q_1, \eta, \nu) = D(Q\|P_1) + \eta\big(D(Q\|P_0) - \beta D(Q\|Q_1) + \gamma(E_0, Q_1)\big) + \nu\left(\sum_{x\in\mathscr{X}} Q(x) - 1\right). \tag{3.15}$$

Theorem 3.3 For every $P_0, P_1 \in \mathscr{P}(\mathscr{X})$ and $E_0 > 0$, we have $\underline{\alpha}_\beta \leq \alpha^*_\beta$ where

$$\underline{\alpha}_\beta = -\Lambda_{\min}(\boldsymbol{H}), \tag{3.16}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

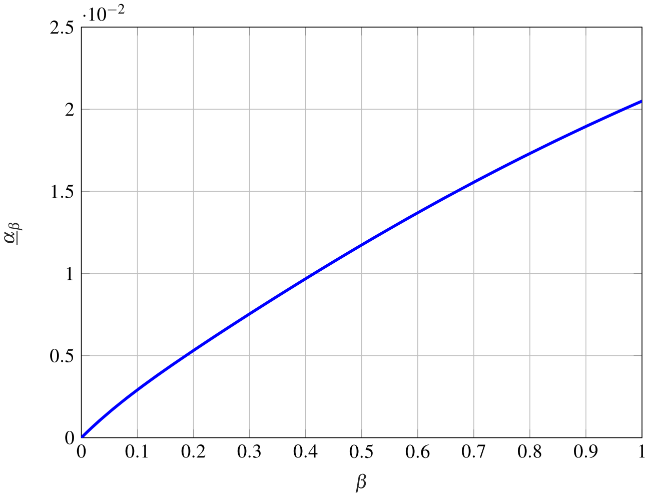

FIG. 4. Lower bound on the required training to test sample ratio $\underline{\alpha}_\beta$ for achieving the optimal error exponent trade-off with the proposed classifier for $0 < \beta \le 1$.

and $\varLambda_{\min}$ is the smallest eigenvalue of the matrix $\boldsymbol{H}$. Moreover, for $|\mathscr{X}| \ge 6$

$$\underline{\alpha}_\beta \ge \beta\eta^*_\beta \left[\frac{Q_{\eta^*_\beta}}{P_1}\right]_{(3)}, \tag{3.17}$$

where $\left[\frac{Q_{\eta^*_\beta}}{P_1}\right]_{(3)}$ is the third largest value of $\frac{Q_{\eta^*_\beta}(i)}{P_1(i)}$ for $i \in \{1, \dots, |\mathscr{X}|\}$.

Proof. The proof can be found in Appendix C. □

EXAMPLE 3.4 Letting the distributions $P_0 = \text{Bern}(0.3), P_1 = \text{Bern}(0.4)$, we have set $E_0 = 0.005$. Figure 4 shows the relation of the $\underline{\alpha}_\beta$ for $0 < \beta < 1$. As can be seen from the figure, $\underline{\alpha}_\beta$ is increasing in β and is equal to zero for $\beta = 0$ as expected, since for $\beta = 0$, the proposed classifier is equal to universal Hoeffding's test and achieves the optimal type-II error exponent for every P_1. However, $\beta = 0$, as discussed before, is the singularity point as the type-I error probability prefactor looses its independence from dimension.

We next find an upper bound on the optimal training to observation ratio α^*_β.

THEOREM 3.5 For every $P_0, P_1 \in \mathscr{P}(\mathscr{X})$ and $E_0 > 0$, we have that $\alpha^*_\beta \le \bar{\alpha}$ where

$$\bar{\alpha} = \frac{\lambda^*(4+\lambda^*)(1+\kappa)}{(P_1^{\min})^2}, \tag{3.18}$$

$P_1^{\min} \triangleq \min_{x\in\mathscr{X}} P_1(x)$, λ^* is the optimal Lagrange multiplier in (2.11) and

$$\kappa = \sqrt{\frac{E_1(E_0, P_1)}{\lambda^*(4+\lambda^*)}}. \tag{3.19}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Proof. The proof can be found in Appendix D. □

Note that the upper bound (3.19) and the lower bound (3.17) to α_β^* suggest that as P_1 approaches the probability simplex boundaries, i.e. as $P_1^{\min}$ approaches zero, the classification problem becomes more challenging, and we need more training samples to achieve the optimal exponents. For the classification problem in Example 3.4, the upper bound in (3.19) gives $\bar{\alpha} = 14.19$, i.e. if the ratio of training samples to the number of test samples exceeds 14.19, the proposed test with $\beta = 1$ achieves (3.6), (3.7). See [22] for general conditions between alphabets size and training sample size when both hypotheses are unknown.

4. Stein regime classification

In this section, we study the classification problem in the Stein regime. For the case where both hypotheses are known, 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 ε (see, e.g. [11])

$$E_1^{(\varepsilon)} \triangleq \sup \left\{ E_1 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \, \varepsilon_0(\phi) \leq \varepsilon \quad \text{and} \quad \varepsilon_1(\phi) \leq e^{-nE_1} \right\}. \tag{4.1}$$

The the optimal $E_1^{(\varepsilon)}$, given by [11]

$$E_1^{(\varepsilon)}(\phi^{\text{lrt}}) = D(P_0 \| P_1), \tag{4.2}$$

can be achieved by setting the threshold in (2.5) to be $\gamma = -D(P_0\|P_1) + \frac{c}{\sqrt{n}}$, where c is a constant that depends on distributions P_0, P_1 and ε. Similarly, our setting where P_0 is known and a training sequence of the second hypothesis is available, we can define the Stein exponent as the highest error exponent under one hypothesis when the error probability under the other hypothesis is at most some fixed ε for all probability distributions $\tilde{P}_1$, i.e.

$$\begin{aligned} E_1^{(\varepsilon)} \triangleq \sup \Big\{ E_1 \in & \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0, \, \forall \tilde{P}_1 \in \mathscr{P}(\mathscr{X}), \\ & \varepsilon_0(\phi | P_0, \tilde{P}_1) \leq \varepsilon \text{ and } \varepsilon_1(\phi | P_0, P_1) \leq e^{-nE_1} \Big\}, \end{aligned} \tag{4.3}$$

where $\varepsilon_0(\phi|P_0, \tilde{P}_1)$ is the error probability when the generating distributions are $P_0, \tilde{P}_1$, i.e. for all possible distributions, $\tilde{P}_1$ the type-I probability of error is bounded by some ε and $E_1^{(\varepsilon)}$ is the maximum achievable type-II error exponent under the actual distribution generating data P_1. Similarly,

$$\begin{aligned} E_0^{(\varepsilon)} \triangleq \sup \Big\{ E_0 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0, \, & \varepsilon_0(\phi | P_0, P_1) \leq e^{-nE_0}, \\ & \varepsilon_1(\phi | P_0, \tilde{P}_1) \leq \varepsilon \;\; \forall \tilde{P}_1 \in \mathscr{P}(\mathscr{X}) \Big\}. \end{aligned} \tag{4.4}$$

In other words, we are interested in the best possible error exponent for one of the hypothesis, while the probability of error under alternative hypothesis is some ε universally for any distribution P_1.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

THEOREM 4.1 Let $\varepsilon \in (0,1)$, then for any probability distributions P_0, P_1, the Stein regime exponents are given by

$$E_1^{(\varepsilon)} = D(P_0 \| P_1), \tag{4.5}$$

$$E_0^{(\varepsilon)} = D_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0), \tag{4.6}$$

where $D_\rho(P_1\|P_0) = \frac{1}{\rho-1}\log\sum_{x\in\mathscr{X}} P_0^\rho P_1^{1-\rho}$ is the Rényi divergence of order ρ.

Proof. The proof can be found in Appendix J. □

Observe that $E_1^{(\varepsilon)}$ is equal to the Stein exponent for the likelihood ratio test where both distributions are known. However, since Rényi divergence is a non-decreasing function of its order and $\frac{\alpha}{1+\alpha} < 1$, hence $E_0^{(\varepsilon)}$ is strictly smaller than the Stein regime exponent achieved by the likelihood ratio test.

5. Sequential classification with a known hypothesis

In this section, we study sequential classification with a known hypothesis. When both hypotheses are known, the SPRT can achieve higher exponents compared with the likelihood ratio test. When only one of the hypotheses is known, Hoeffding's test can achieve the best error exponent achieved by the likelihood ratio test in the fixed sample size scenario. However, there is no counterpart to Hoeffding's test in the sequential case when only one of the hypotheses is known, i.e. no classifier can achieve the same error exponent performance as the SPRT [39]. We propose a classifier inspired by the SPRT and show that having training samples from the second hypothesis can improve the error exponent tradeoff compared with the fixed sample-sized classification.

In the sequential setting, the number of samples is a random variable called the stopping time τ, taking values in $\mathbb{Z}_+$. A sequential classifier is a pair $\Phi = (\phi : \mathscr{X}^\tau \times \mathscr{X}^{\alpha\tau} \to \{0,1\}, \tau)$, where for every $n \geq 0$, the event $\{\tau \leq n\} \in \mathscr{F}_n$, and $\mathscr{F}_n$ is the sigma-algebra induced by random variables $\boldsymbol{x}^n, \boldsymbol{z}^{\alpha n}$, i.e. $\sigma(\boldsymbol{x}^n, \boldsymbol{z}^{\alpha n})$. We also assume that at every stage, additional training and test samples are available to the classifier, such that $\alpha = \frac{k}{n}$ remains constant. Moreover, ϕ is a $\mathscr{F}_\tau$ measurable decision rule, i.e. the decision rule determined by casually observing the sequence $\boldsymbol{x}^n, \boldsymbol{z}^{\alpha n}$. In other words, at each time instant, the test attempts to decide in favor of one of the hypotheses or chooses to take new samples from the source P_1 as well as new samples from the unknown data source.

The two possible pairwise error probabilities that measure the performance of the test are defined as

$$\varepsilon_0(\Phi) = \mathbb{P}_0\big[\phi(\boldsymbol{x}^\tau, \boldsymbol{z}^{\alpha\tau}) \neq 0\big], \varepsilon_1(\Phi) = \mathbb{P}_1\big[\phi(\boldsymbol{x}^\tau, \boldsymbol{z}^{\alpha\tau}) \neq 1\big], \tag{5.1}$$

where the probabilities are over P_0, P_1, respectively. Similarly to the sequential hypothesis testing case, we define the optimal error exponent as

$$\begin{aligned} E_1^*(E_0) \triangleq \sup \Big\{ & E_1 \in \mathbb{R}^+ : \exists \Phi,\ \exists n' \in \mathbb{Z}_+ \text{ s.t. } \forall n_0 > n', n_1 > n', \\ & \mathbb{E}_{P_0}[\tau] \leq n_0, \mathbb{E}_{P_1}[\tau] \leq n_1,\ \varepsilon_0(\Phi) \leq 2^{-n_0 E_0} \text{ and } \varepsilon_1(\Phi) \leq 2^{-n_0 E_1} \Big\}. \end{aligned} \tag{5.2}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

When both hypotheses are known, the SPRT $\Phi = (\phi, \tau)$ proposed by Wald [38] achieves the optimal exponent tradeoff. The SPRT is given by

$$\tau = \inf\left\{t \geq 1 : S_t \geq \gamma_0 \text{ or } S_t \leq -\gamma_1\right\}, \tag{5.3}$$

where

$$S_n = \sum_{i=1}^{t} \log \frac{P_0(x_i)}{P_1(x_i)}, \tag{5.4}$$

is the accumulated log-likelihood ratio (LLR) of the observed sequence $\boldsymbol{x}$ and the thresholds γ_0, γ_1 are two positive real numbers. Moreover, the test makes a decision according to the rule

$$\phi(\hat{T}_{\boldsymbol{x}}) = \begin{cases} 0 & \text{if } S_\tau \geq \gamma_0 \\ 1 & \text{if } S_\tau \leq -\gamma_1. \end{cases} \tag{5.5}$$

It is shown in [38] that the above test attains the optimal error exponent tradeoff, i.e., as thresholds γ_0, γ_1 approach infinity, the test achieves the best error exponent trade-off in (5.2). It is known that the error probabilities of SPRT as a function of γ_0 and γ_1 are [40]

$$\varepsilon_0 = c_0 \cdot e^{-\gamma_1} \quad , \quad \varepsilon_1 = c_1 \cdot e^{-\gamma_0}, \tag{5.6}$$

as $\gamma_0, \gamma_1 \to \infty$, where c_0, c_1 are positive constants. Moreover, it can also be shown that

$$\mathbb{E}_{P_0}[\tau] = \frac{\gamma_0}{D(P_0\|P_1)}(1 + o(1)), \tag{5.7}$$

$$\mathbb{E}_{P_1}[\tau] = \frac{\gamma_1}{D(P_1\|P_0)}(1 + o(1)). \tag{5.8}$$

Therefore, according to definition (5.2), the optimal error exponent tradeoff is given by

$$E_0 \cdot E_1 \leq D(P_1\|P_0) \cdot D(P_0\|P_1), \tag{5.9}$$

where thresholds γ_0, γ_1 are chosen as

$$\gamma_0 = n_0\big(D(P_0\|P_1) + o(1)\big), \ \gamma_1 = n_1\big(D(P_1\|P_0) + o(1)\big). \tag{5.10}$$

Hence, the SPRT achieves the Stein regime error exponents achievable by the standard likelihood ratio test [30] simultaneously.

For every fixed n. we propose the following sequential classifier:

$$\tau = \inf\left\{t \geq n : S_t(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) \geq \gamma_{0,n}(t) \text{ or } S_t(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) \leq -\gamma_{1,n}(t)\right\}, \tag{5.11}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where

$$S_t(\hat{T}_{\boldsymbol{x}}, \hat{T}_z) = \sum_{i=1}^{t} \log \frac{P_0(x_i)}{\hat{T}'_z(x_i)}, \tag{5.12}$$

is the accumulated LLR using the plugin perturbed type of the training sequence evaluated at the observed sequence $\boldsymbol{x}$, and

$$\hat{T}'_z = (1 - \delta_n)\hat{T}_z + \frac{\delta_n}{|\mathscr{X}|}, \tag{5.13}$$

where $\delta_n = o(n^{-1})$ and $\gamma_{0,n}(t), \gamma_{1,n}(t)$ are chosen as

$$\gamma_{0,n}(t) = nD(P_0\|\hat{T}'_z) + (4|\mathscr{X}| + 4)\log(t+1),\ \gamma_{1,n}(t) = nD(\hat{T}_z\|P_0) + (4|\mathscr{X}| + 4)\log(t+1), \tag{5.14}$$

and the test makes a decision according to the rule

$$\phi(\hat{T}_{\boldsymbol{x}}, \hat{T}_z) = \begin{cases} 0 & \text{if } S_\tau(\hat{T}_{\boldsymbol{x}}, \hat{T}_z) \geq \gamma_{0,n}(t) \\ 1 & \text{if } S_\tau(\hat{T}_{\boldsymbol{x}}, \hat{T}_z) \leq -\gamma_{1,n}(t), \end{cases} \tag{5.15}$$

where $\hat{T}_{\boldsymbol{x}}, \hat{T}_z$ are types of the test and training samples at the stopping time τ. As can be seen from the above expressions, the proposed classifier is the plugin SPRT, replacing P_1 by the perturbed training type $\hat{T}'_z$.

The next theorem gives a lower bound on the achievable error exponent tradeoff of the proposed sequential classifier.

Theorem 5.1 For every P_0, P_1, there exists a training to observation ratio α^*_{seq} such that for any $\alpha \geq \alpha^*_{\mathrm{seq}}$, the sequential classifier $\Phi^{\mathrm{seq}} = (\phi(\hat{T}_{\boldsymbol{x}}, \hat{T}'_z), \tau)$ defined in (5.11), (5.15) achieves

$$E_0(\Phi^{\mathrm{seq}})E_1(\Phi^{\mathrm{seq}}) \geq D(P_0\|P_1)D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0). \tag{5.16}$$

Furthermore, the average stopping times of the classifier satisfy

$$\mathbb{E}_{P_0}[\tau] = \mathbb{E}_{P_1}[\tau] = n(1 + o(1)). \tag{5.17}$$

Proof. The proof can be found in Appendix K. □

This theorem shows that similar to the hypothesis testing problem with known distributions, the proposed sequential classifier can achieve the Stein regime exponents simultaneously when only one of the distributions is known.

Example 5.2 In Fig. 5, we present a numerical example to illustrate the performance of the proposed sequential classifier in Φ^{seq} with $\gamma_{0,n}(t) = nD(P_0\|\hat{T}'_z), \gamma_{1,n}(t) = nD(\hat{T}_z\|P_0)$. Consider two binary distributions $P_0 = Bern(0.45)$, $Bern(0.55)$ and set $\alpha = 10$. Each point in the figures 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 sequential classifier in (5.12) by generating a sample

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

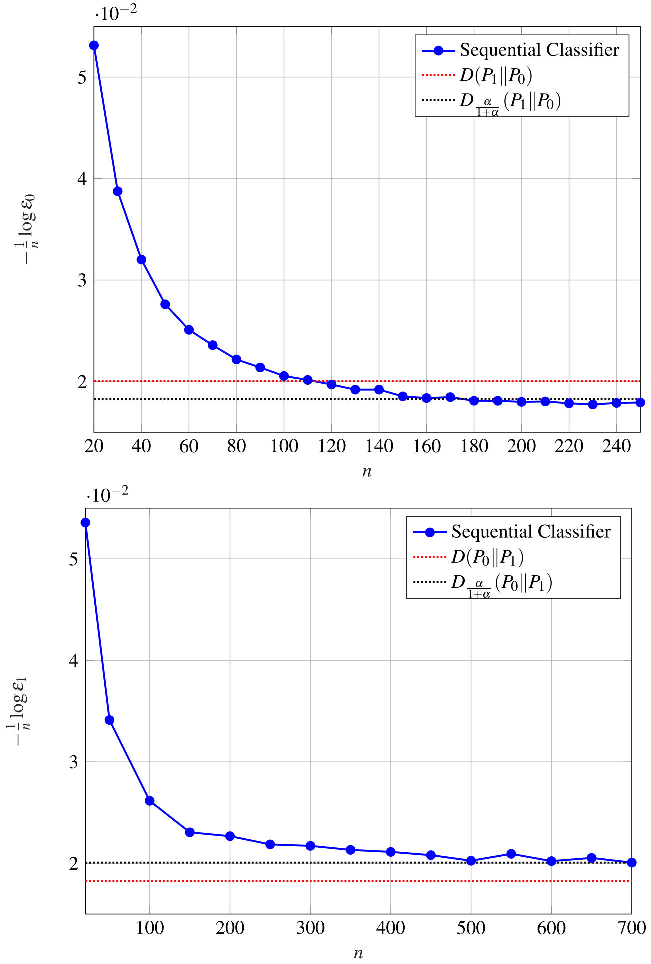

FIG. 5. Type-I and type-II error exponents for the proposed sequential classifier.

from the test source and α samples from P_1 until the test stops and makes a decision. We have plotted the $-\frac{1}{n}\log\varepsilon_i$ for $i \in \{0,1\}$. We can notice that the type-I error exponent converges to $D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$, while the type-II error exponent tends to $D(P_0\|P_1)$ as Theorem 5.1 suggests.

THEOREM 5.3 For every sequential classifier $\Phi^{\text{seq}} = (\phi^{\text{seq}}(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}), \tau)$, such that $\mathbb{E}_{P_0}[\tau] \leq n, \mathbb{E}_{P_1}[\tau] \leq n$, we have

$$\max_{\Phi^{\text{seq}}} \min_{P_1 \in \mathscr{P}(\mathscr{X})} E_0(\Phi^{\text{seq}})E_1(\Phi^{\text{seq}}) \leq D(P_0\|P_1)D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0). \tag{5.18}$$

This suggests that the proposed sequential classifier is universal in the sense that it achieves the highest error exponent tradeoff over all possible classifiers and distributions P_1.

Proof. The proof can be found in Appendix L. □

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

6. Conclusions

We have proposed and analyzed a new universal classifier where the null hypothesis is known and only training samples are available for the alternative hypothesis. The new classifier combines the log-likelihood test as well as Hoeffding's test, and not only attains the optimal error exponent tradeoff but also the optimal prefactor order provided that the training sequence is sufficiently long. We have proposed the sequential version of the classifier, which is attains the optimal error exponent tradeoff in the sequential setting.

Acknowledgements

We would like to thank the anonymous reviewers for the thorough review and detailed comments, which have significantly improved the paper.

Funding

European Research Council (725411); Spanish Ministry of Economy and Competitiveness (PID2020-116683GB-C22).

Data availability

No new data were generated or analysed in support of this research.

1. Berge, C. (1997) *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Dover Publications: Dover Books on Mathematics.
2. Billingsley, P. (1986) *Probability and Measure*, 2nd edn. New York, NY: John Wiley and Sons.
3. Boroumand, P. & Guillén i Fàbregas, A. (2022) Mismatched binary hypothesis testing: error exponent sensitivity. *IEEE Trans. Inf. Theory*, **68**, 6738–6731, 6761.
4. Boucheron, S., Lugosi, G. & Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press.
5. Boyd, S. & Vandenberghe, L. (2004) *Convex Optimization*. Cambridge, UK: Cambridge University Press.
6. Bu, Y., Zou, S., Liang, Y. & Veeravalli, V. (2018) Estimation of KL divergence: optimal minimax rate. *IEEE Trans. Inf. Theory*, **64**(4), 2648–2674.
7. Castillo, E., Conejo, A. J., Castillo, C., Mínguez, R. & Ortigosa, D. (2006) Perturbation approach to sensitivity analysis in mathematical programming. *J. Optim. Theory Appl.*, **128**(1), 49–74.
8. Chakravarti, I. (1980) *Asymptotic Theory of Statistical Tests and Estimation: In Honor of Wassily Hoeffding*. Cambridge, MA, USA: Academic Press.
9. Cortes, C. & Vapnik, V. (1995) *Support-vector networks*. Mach. Learn., **20**, 273–297.
10. Cover, T. & Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, **13**, 21–27.
11. Cover, T. M. & Thomas, J. A. (2006) *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Hoboken, NJ, USA: Wiley-Interscience.
12. Csiszár, I. & Körner, J. (2011) *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, UK: Cambridge University Press.
13. Csiszár, I. & Shields, P. (2004) Information theory and statistics: a tutorial. *Found. Trends Commun. Inf. Theory*, **1**, 417–528.
14. Dembo, A. & Zeitouni, O. (2010) *Large Deviations Techniques and Applications*, vol. **38**. New York, NY, USA: Springer.
15. Devroye, L., Györfi, L. & Lugosi, G. (2013) *A probabilistic theory of pattern recognition*, vol. **31**. New York, NY, USA: Springer Science & Business Media.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

16. Gutman, M. (1989) Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Inf. Theory*, **35**, 401–408.

17. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. **2**. New York, NY, USA: Springer.

18. Hoeffding, W. (1994) *Asymptotically Optimal Tests for Multinomial Distributionspp*. New York, NY: Springer, pp. 431–471.

19. Horn, R. A. & Johnson, C. R. (1991) *Topics in Matrix Analysis*. Cambridge, UK: Cambridge University Press.

20. Iltis, M. (1995) Sharp asymptotics of large deviations. *J. Theoret. Probab.*, **8**, 501–522.

21. Karson, M. (1968) Handbook of methods of applied statistics. Volume I: techniques of computation descriptive methods, and statistical inference. Volume II: planning of surveys and experiments. I. M. Chakravarti, R. G. Laha, and J. Roy, New York, John Wiley; 1967, $9.00. *J. Amer. Statist. Assoc.*, **63**, 1047–1049.

22. Kelly, B. G., Wagner, A. B., Tularak, T. & Viswanath, P. (2013) Classification of homogeneous data with large alphabets. *IEEE Trans. Inf. Theory*, **59**, 782–795.

23. Lee, J. M. (2013) Smooth manifolds. In: *Introduction to Smooth Manifolds*. New York, NY, USA: Springer, pp. 1–31.

24. Lehmann, E. L. & Romano, J. P. (2005) *Testing statistical hypotheses, Springer Texts in Statistics*, 3rd edn. New York: Springer.

25. Li, Y., Nitinawarat, S. & Veeravalli, V. V. (2014) Universal outlier hypothesis testing. *IEEE Trans. Inf. Theory*, **60**, 4066–4082.

26. Li, Y. & Tan, V. Y. F. (2020) Second-order Asymptotics of sequential hypothesis testing. *IEEE Trans. Inf. Theory*, **66**, 7222–7230.

27. Milgrom, P. & Segal, I. (2002) Envelope theorems for arbitrary choice sets. *Econometrica*, **70**, 583–601.

28. Moulin, P. (2017) The log-volume of optimal codes for memoryless channels, asymptotically within a few nats. *IEEE Trans. Inf. Theory*, **63**, 2278–2313.

29. Ney, P. (1983) Dominating Points and the Asymptotics of Large Deviations for Random Walk on $\mathbb{R}^d$. *Ann. Probab.*, **11**, 158–167.

30. Neyman, J. & Pearson, E. S. (1933) 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*, **231**, 289–337.

31. Polyanskiy, Y., Poor, H. V. & Verdú, S. (2010) Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, **56**, 2307–2359.

32. Polyanskiy, Y. & Verdú, S. (2011) Binary hypothesis testing with feedback. *Information Theory and Applications Workshop (ITA)*. CA, USA: San Diego.

33. Resnick, S. (2019) *A probability Path*. New York, NY, USA: Springer.

34. Rudin, W. (1964) *Principles of mathematical analysis*, vol. **3**. New York: McGraw-hill.

35. Tong, X. (2013) A plug-in approach to Neyman-Pearson classification. *J. Mach. Learn. Res.*, **14**, 3011–3040.

36. van Erven, T. & Harremoës, P. (2014) Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*, **60**, 3797–3820.

37. Vazquez-Vilar, G., Guillén i Fàbregas, A., Koch, T. & Lancho, A. (2018) Saddlepoint approximation of the error probability of binary hypothesis testing. *2018 IEEE International Symposium on Information Theory (ISIT)*. CO, USA: Vail, pp. 2306–2310.

38. Wald, A. (1945) Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, **16**, 117–186.

39. Wald, A. & Wolfowitz, J. (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, **19**, 326–339.

40. Woodroofe, M. (1982) *Nonlinear Renewal Theory in Sequential Analysis*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

41. Zeitouni, O. & Gutman, M. (1991) On universal hypotheses testing via large deviations. *IEEE Trans. Inf. Theory*, **37**, 285–290.

42. Zhou, L., Tan, V. Y. F. & Motani, M. (2019) Second-order asymptotically optimal statistical classification. *Inf. Inference*, **9**, 81–111.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

A. Supporting results

The proofs of the main results rely extensively on the method of types [12]. The type of an n-length sequence $\boldsymbol{y}$ is defined as $\hat{T}_{\boldsymbol{y}}(a) = \frac{N(a|\boldsymbol{y})}{n}$, where $N(a|\boldsymbol{y})$ is the number of occurrences of symbol $a \in \mathcal{X}$ in sequence $\boldsymbol{y}$. The set of all sequences of length n with type P, denoted by $\mathcal{T}_P^n$, is called the type class. The set of types with length n sequences on the simplex $\mathcal{P}(\mathcal{X})$ is denoted as $\mathcal{P}_n(\mathcal{X})$. It is well known that $|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|-1}$ and that for a type P and probability distribution Q

$$Q(\mathcal{T}_P^n) = \sum_{\boldsymbol{x} \in \mathcal{T}_P^n} Q(\boldsymbol{x}) \leq e^{-nD(P\|Q)}. \tag{A.1}$$

A refinement of the above is given by [12, Ch. 2]

$$Q(\mathcal{T}_Q^n) = n^{\frac{-|\mathcal{X}|+1}{2}} e^{-nD(P\|Q)}(c + o(1)). \tag{A.2}$$

We use the shorthand notation $a_n \asymp b_n$ for any two positive real sequences such that $\log \frac{a_n}{b_n} = \mathcal{O}(1)$. The following result is used to asymptotically convert the summations we obtain from the method of types into integrals, so we can then apply Laplace's integration method.

Theorem A.1 [8, Theorem 1] 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,

$$\sum_{\boldsymbol{w} \in \mathcal{D} \cap \mathcal{P}_k(\mathcal{X})} e^{-k\psi(\boldsymbol{w})} \asymp k^d \int_{\mathcal{D}} e^{-k\psi(\boldsymbol{w})}\, d\boldsymbol{w}. \tag{A.3}$$

The precise definition of minimally smooth boundary can be found in [8]. In this paper, we only work with smooth boundaries which are also minimally smooth

Next we introduce the following theorem which can be derived from [20]. The theorem uses Laplace's integration method to expand the integral exponent around the dominating point and getting the prefactor by careful expansion of the terms. We use it to approximate the integral on the right hand side of (A.3).

Theorem A.2 [20, Theorem 1.4] For a set Γ in $\mathbb{R}^d$ having smooth boundary with a unique dominating point ν where the large deviation rate function $I(x)$ is minimized with $I(\nu) = a$, let the level surface $S_a = \{x \in \mathbb{R}^d : I(x) = a\}$ be described locally by $x_d = f(x)$, where $x = (x_1, \ldots, x_{d-1})$, and let the region Γ be bounded by the surface $x_d = g(x)$, where $g(x)$ is a three times differentiable function of x in a neighbourhood of 0. If the Hessians of g and f at 0 satisfy $H_g(0) > H_s(0)$, then

$$\int_{\boldsymbol{w} \in \Gamma} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w} = n^{-\frac{d+1}{2}} e^{-na}(d_0 + o(1)). \tag{A.4}$$

B. Proof of Theorem 3.1

First, we prove that for every realization of the training sequence $\hat{T}_z$, the type-I error exponent is equal to E_0. Note that the solution to the optimization problem in (3.2) is a convex problem, and by the Karush–Kuhn–Tucker (KKT) conditions [5], the minimizer is unique and is the tilted distribution that interpolates

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

between P_0 and $\hat{T}_{z'}$, i.e.

$$Q_{\mu^*}(x) = \frac{P_0^{\frac{\mu^*}{1+\mu^*}}(x)\hat{T}_{z'}^{\frac{1}{1+\mu^*}}(x)}{\sum_{a\in\mathscr{X}} P_0^{\frac{\mu^*}{1+\mu^*}}(a)\hat{T}_{z'}^{\frac{1}{1+\mu^*}}(a)}, \tag{B.1}$$

and μ^* is the solution to

$$D(Q_{\mu^*}\|P_0) = E_0. \tag{B.2}$$

Therefore, the classifier threshold in (3.2) can be written as

$$\gamma(E_0, \hat{T}'_z) = \beta D(Q_{\mu^*}\|\hat{T}'_z) - D(Q_{\mu^*}\|P_0). \tag{B.3}$$

Now by Sanov's theorem, we can find the type-I error exponent by solving (2.7) with P_1 replaced by $\hat{T}_{z'}$ and γ replaced by (B.3). This optimization problem is convex in Q for every $\hat{T}_{z'}$ when $\beta = 1$, however for $\beta < 1$, this is not the case and the KKT conditions are only necessary conditions. Spelling out the Lagrangian, we obtain the KKT conditions

$$L(Q,\lambda,\nu) = D(Q\|P_0) + \lambda\big(\beta D(Q\|\hat{T}'_z) - D(Q\|P_0) - \gamma(E_0,\hat{T}'_z)\big) + \nu\left(\sum_{x\in\mathscr{X}} Q(x) - 1\right) \tag{B.4}$$

$$\frac{\partial L(Q,\lambda,\nu)}{\partial Q(x)} = 1 + \log\frac{Q(x)}{P_0(x)} + \lambda\left(\beta + \beta\log\frac{Q(x)}{\hat{T}'_z(x)} - 1 - \log\frac{Q(x)}{P_0(x)}\right) + \nu. \tag{B.5}$$

Setting the derivative to zero, we get

$$Q_{\beta,\lambda^*}(x) = \frac{P_0^{\frac{1-\lambda^*}{1-\lambda^*+\lambda^*\beta}}(x)\hat{T}_{z'}^{\frac{\lambda^*\beta}{1-\lambda^*+\lambda^*\beta}}(x)}{\sum_{a\in\mathscr{X}} P_0^{\frac{1-\lambda^*}{1-\lambda^*+\lambda^*\beta}}(a)\hat{T}_{z'}^{\frac{\lambda^*\beta}{1-\lambda^*+\lambda^*\beta}}(a)}, \quad 0 \le \lambda^*, \tag{B.6}$$

and by complementary slackness condition [5]

$$D(Q_{\beta,\lambda^*}\|P_0) - \beta D(Q_{\beta,\lambda^*}\|\hat{T}'_z) = D(Q_{\mu^*}\|P_0) - \beta D(Q_{\mu^*}\|\hat{T}'_z). \tag{B.7}$$

Note that, λ^* cannot be zero as that sets $Q_{\beta,\lambda^*} = P_0(x)$ which is invalid solution when $E_0 > 0$. Observe that this equality is satisfied when $\frac{1-\lambda^*}{1-\lambda^*+\lambda^*\beta} = \frac{\mu^*}{1+\mu^*}$. This is the unique solution as $\frac{1-\lambda^*}{1-\lambda^*+\lambda^*\beta}$ is strictly decreasing function in λ^*, and hence, $D(Q_{\beta,\lambda^*}\|P_0)$ and $D(Q_{\beta,\lambda^*}\|\hat{T}_{z'})$ are strictly increasing and strictly decreasing in λ^*, respectively. Therefore, we get

$$D(Q_{\beta,\lambda^*}\|P_0) = E_0. \tag{B.8}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Next, by rewriting the type-I error probability as a function of the types of the training and observation sequences $\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}$, we have

$$\varepsilon_0(\phi_\beta) = \sum_{\hat{T}_{\boldsymbol{z}} \in \mathscr{P}_k(\mathscr{X})} P_1(\hat{T}_{\boldsymbol{z}}) \sum_{\substack{\hat{T}_{\boldsymbol{x}} \in \mathscr{P}_n(\mathscr{X}), \\ \phi_\beta(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}})=1}} P_0(\hat{T}_{\boldsymbol{x}}), \tag{B.9}$$

that is the type-I error exponent is equal to E_0 for every realization of training sequence $\hat{T}_{\boldsymbol{z}}$. Since the hypothesis test is a type-based test, and using (A.2), we obtain

$$\varepsilon_0(\phi_\beta) = \sum_{\hat{T}_{\boldsymbol{z}} \in \mathscr{P}_k(\mathscr{X})} P_1(\hat{T}_{\boldsymbol{z}}) \sum_{\hat{T}_{\boldsymbol{x}} \in \mathscr{A}_1(\hat{T}_{\boldsymbol{z}}, \beta)} n^{\frac{-\left|\mathscr{X}\right|+1}{2}} e^{-nD(\hat{T}_{\boldsymbol{x}} \| P_0)}(c + o(1)). \tag{B.10}$$

In order to find the polynomial decay of the error probability, we use the Theorems A.1 and A.2 [8, 20, 29], to approximate the summation by an integral and then use Laplace's integration method.

For a fixed $\hat{T}_{\boldsymbol{z}'}$, it can be shown that $\phi(Q) = D(Q\|P_0) - \beta D(Q\|\hat{T}_{\boldsymbol{z}'})$ is a smooth function, and hence, we can conclude that the boundary of the decision region $\mathscr{A}_0(\hat{T}_{\boldsymbol{z}}, \beta)$ is smooth since the graph of a smooth function is a smooth manifold [23]. Furthermore, by the continuous differentiability of $\psi(Q) = D(Q\|P_0)$ in Q, the conditions of the Theorem A.1 are satisfied and we can approximate the summation in (B.10) by the integral

$$\sum_{\boldsymbol{w} \in \mathscr{A}_1(\hat{T}_{\boldsymbol{z}}, \beta) \cap \mathscr{P}_n(\mathscr{X})} e^{-nD(\boldsymbol{w}\|P_0)} \asymp n^{\left|\mathscr{X}\right|-1} \int_{\mathscr{A}_1(\hat{T}_{\boldsymbol{z}}, \beta)} e^{-nD(\boldsymbol{w}\|P_0)} d\boldsymbol{w}. \tag{B.11}$$

Since the optimizing distribution solving Sanov's Theorem is unique and equal to Q_{λ^*}, the dominating point is unique. Hence, we only need to prove that the Hessian of the decision region $\mathscr{D}$ minus the Hessian of the level surface $\mathscr{S}_{E_0} = \{Q \in \mathscr{P}(\mathscr{X}) : D(Q\|P_0) \le E_0\}$ is a positive definite matrix. Writing $Q_{\beta,\lambda^*}(|\mathscr{X}|) = 1 - \sum_{x=1}^{|\mathscr{X}|-1} Q_{\beta,\lambda^*}(x)$ and taking the derivatives, we have

$$\boldsymbol{H}_{\mathscr{D}}^{(|\mathscr{X}|-1)\times(|\mathscr{X}|-1)} = -(1-\beta)\left(\operatorname{diag}\left(\frac{1}{Q_{\beta,\lambda^*}}\right) + \frac{\mathbf{1}\mathbf{1}^T}{Q_{\beta,\lambda^*}(|\mathscr{X}|)}\right), \tag{B.12}$$

$$\boldsymbol{H}_{\mathscr{S}_{E_0}}^{(|\mathscr{X}|-1)\times(\left|\mathscr{X}\right|-1)} = -\operatorname{diag}\left(\frac{1}{Q_{\beta,\lambda^*}}\right) - \frac{\mathbf{1}\mathbf{1}^T}{Q_{\beta,\lambda^*}(|\mathscr{X}|)}, \tag{B.13}$$

and hence

$$\boldsymbol{H}_{\mathscr{S}_{E_0}} - \boldsymbol{H}_{\mathscr{D}} = \beta \operatorname{diag}\left(\frac{1}{Q_{\beta,\lambda^*}}\right) + \beta \frac{\mathbf{1}\mathbf{1}^T}{Q_{\beta,\lambda^*}(|\mathscr{X}|)}, \tag{B.14}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

which is positive for any $\beta > 0$. Therefore, by (B.10), (B.11), (A.4), we have

$$\varepsilon_0(\phi_\beta) = \sum_{\hat{T}_z} P_1(\hat{T}_z) n^{\frac{-|\mathscr{X}|+1}{2}} n^{|\mathscr{X}|-1} n^{-\frac{|\mathscr{X}|}{2}} e^{-nE_0}(c + o(1)) \tag{B.15}$$

$$= \sum_{\hat{T}_z} P_1(\hat{T}_z) \frac{1}{\sqrt{n}} e^{-nE_0}(c + o(1)). \tag{B.16}$$

Finally, since c is finite, positive and only depends on P_0, P_1, E_0, there exists a $\tilde{c}$ such that

$$\varepsilon_0(\phi_\beta) = \frac{1}{\sqrt{n}} e^{-nE_0}\big(\tilde{c} + o(1)\big), \tag{B.17}$$

which completes the first statement.

Having established the improvement of the prefactor of the type-I error probability, we are ready to study the type-II error exponent of our proposed classifier. Using standard properties of the method of types [12], the type-II probability of error can be written as

$$\varepsilon_1(\phi_\beta) = \sum_{\boldsymbol{x},\boldsymbol{z}:\phi_\beta(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})=0} P_1(\boldsymbol{z}) P_1(\boldsymbol{x}) \tag{B.18}$$

$$= \sum_{\substack{Q \in \mathscr{P}_n(\mathscr{X}), Q_1 \in \mathscr{P}_k(\mathscr{X}), \\ \phi_\beta(Q,Q_1)=0}} P_1\big(\mathscr{T}_Q^n\big) P_1\big(\mathscr{T}_{Q_1}^k\big) \tag{B.19}$$

$$\leq \sum_{\substack{Q \in \mathscr{P}_n(\mathscr{X}), Q_1 \in \mathscr{P}_k(\mathscr{X}), \\ \phi_\beta(Q,Q_1)=0}} e^{-nD(Q\|P_1) - kD(Q_1\|P_1)} \tag{B.20}$$

$$\leq (n+1)^{|\mathscr{X}|}(k+1)^{|\mathscr{X}|} 2^{-n\tilde{E}_1^{(n)}(\phi_\beta)}, \tag{B.21}$$

where

$$\tilde{E}_1^{(n)}(\phi_\beta) = \min_{\substack{\phi_\beta(Q,Q_1)=0 \\ Q,Q_1 \in \mathscr{P}(\mathscr{X})}} D(Q\|P_1) + \alpha D(Q_1\|P_1). \tag{B.22}$$

It is worth observing that this optimization problem is non-convex. In addition, lower bounding (B.19) yields

$$\varepsilon_1(\phi_\beta) \geq (n+1)^{-|\mathscr{X}|}(k+1)^{-|\mathscr{X}|} e^{-n\tilde{E}_1^{(n)}(\phi_\beta)}. \tag{B.23}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Observe that the proposed test ϕ_β depends on n; hence, if the limit of (B.22) as n tends to infinity exists, then the limit is the type-II error exponent. Using a change of variable, we can write (B.22) as

$$\begin{aligned} \tilde{E}_1^{(n)}(\phi_\beta) = \min_{Q,Q_1} & D(Q\|P_1) + \alpha D\left(\frac{1}{1-\delta_n}Q_1 - \frac{\delta_n}{1-\delta_n}U\middle\|P_1\right) \\ \text{s.t. } & \beta D(Q\|Q_1) - D(Q\|P_0) \geq \gamma(E_0, Q_1) \\ & Q \in \mathscr{P}(\mathscr{X}) \\ & Q_1 \in \big(1-\delta_n\big)\mathscr{P}(\mathscr{X}) + \delta_n U, \end{aligned} \tag{B.24}$$

where U is the equiprobable distribution on $\mathscr{P}(\mathscr{X})$. Now, by using a Taylor expansion of $D\big(\frac{1}{1-\delta_n}Q_1 - \frac{\delta_n}{1-\delta_n}U\|P_1\big)$ around $\delta_n = 0$, we get

$$\begin{aligned} \tilde{E}_1^{(n)}(\phi_\beta) = \min_{Q,Q_1} \quad & D(Q\|P_1) + \alpha D(Q_1\|P_1) + \mathscr{O}\left(\sum_{x\in\mathscr{X}} \log\frac{Q_1(x)}{P_1(x)}\frac{\delta_n}{1-\delta_n}\right) \\ \text{s.t.} \quad & \beta D(Q\|Q_1) - D(Q\|P_0) \geq \gamma(E_0, Q_1) \\ & Q \in \mathscr{P}(\mathscr{X}) \\ & Q_1 \in (1-\delta_n)\mathscr{P}(\mathscr{X}) + \delta_n U. \end{aligned} \tag{B.25}$$

Note that $Q_1(x) \leq 1$ as $Q_1(x)$ is a probability distribution, and $P_1(x) > 0$ for all $x \in \mathscr{X}$ by assumption. Hence, the remainder term can be dropped. Therefore, we can approximate the type-II error exponent by

$$E_1^{(n)}(\phi_\beta) = \min_{\substack{\beta D(Q\|Q_1) - D(Q\|P_0) \geq \gamma(E_0,Q_1) \\ Q\in\mathscr{P}(\mathscr{X}), Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} D(Q\|P_1) + \alpha D(Q_1\|P_1), \tag{B.26}$$

where

$$\mathscr{P}_{\delta_n}(\mathscr{X}) = \big\{Q : Q = (1-\delta_n)P + \delta_n U, P \in \mathscr{P}(\mathscr{X})\big\}, \tag{B.27}$$

and the approximation error is of order $\mathscr{O}\big(\frac{\delta_n}{1-\delta_n}\big)$. Letting $\delta_n = o\big(n^{-1}\big)$, we have

$$\varepsilon_1(\phi_\beta) = e^{-nE_1^{(n)}(\phi_\beta)+o(1)} \asymp e^{-nE_1^{(n)}(\phi_\beta)}. \tag{B.28}$$

Hence, the type-II error exponent can be calculated as the limit of (B.26) when n tends to infinity. We also define the following optimization problems. For every $Q_1 \in \mathscr{P}_{\delta_n}(\mathscr{X})$, let

$$E_1(E_0, Q_1) = \min_{\substack{\beta D(Q\|Q_1) - D(Q\|P_0) \geq \gamma(E_0,Q_1) \\ Q\in\mathscr{P}(\mathscr{X}),}} D(Q\|P_1), \tag{B.29}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

which is the error exponent when the type of the training sequence is Q_1. Also, let

$$E_1^{(n)}(E_0, r) = \min_{\substack{D(Q_1\|P_1)\leq r\\ Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} E_1(E_0, Q_1). \tag{B.30}$$

The latter optimization problem is the worst case achievable error exponent by ϕ_β if we know the training sequence type Q_1 is inside a relative entropy ball around the original distribution P_1 of radius r [3]. Using (B.29), (B.30), we can write (B.26) as

$$\begin{aligned}
E_1^{(n)}(\phi_\beta) &= \min_{\substack{\beta D(Q\|Q_1)-D(Q\|P_0)\geq\gamma(E_0,Q_1),\\ D(Q_1\|P_1)=r,\, r\geq 0,\\ Q\in\mathscr{P}(\mathscr{X}),\, Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} D(Q\|P_1) + \alpha r & \text{(B.31)}\\
&= \min_{\substack{D(Q_1\|P_1)=r,\, r\geq 0,\\ Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} E_1(E_0, Q_1) + \alpha r & \text{(B.32)}\\
&= \min_{\substack{D(Q_1\|P_1)\leq r,\, r\geq 0\\ Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} E_1(E_0, Q_1) + \alpha r & \text{(B.33)}\\
&= \min_{r\geq 0} E_1^{(n)}(E_0, r) + \alpha r, & \text{(B.34)}
\end{aligned}$$

where in (B.33) we used the KKT conditions to show that the minimizer Q_1^* of (B.33) should satisfy the inequality condition $D(Q_1\|P_1) \leq r$ with equality, i.e. $D(Q_1^*\|P_1) = r$. We will use (B.34) to analyze the behaviour of the non-convex problem in (B.22). We also define r_c as

$$r_c = \min_{\substack{Q_1\in\mathscr{P}(\mathscr{X})\\ \min_{i\in\mathscr{X}} Q_1(i)=0}} D(Q_1\|P_1), \tag{B.35}$$

that is the minimum distance between P_1 and distributions on the boundary of the probability simplex $\mathscr{P}(\mathscr{X})$. Since $P_1(i)$ is non-zero for every i, therefore $r_c > 0$, and we can lower bound $E_1^{(n)}(\phi_\beta)$ by

$$E_1^{(n)}(\phi_\beta) \geq \min\left\{ \min_{0\leq r\leq r_c-\varepsilon} E_1^{(n)}(E_0, r) + \alpha r, \alpha(r_c - \varepsilon)\right\}, \tag{B.36}$$

where $0 < \varepsilon < r_c$, and we used the fact that $E_1^{(n)}(E_0, r) \geq 0$ for $r \geq 0$ by the non-negativity of the relative entropy. For $r \in [0, r_c - \varepsilon]$, we also define $E_1(E_0, r)$ as

$$\begin{aligned}
E_1(E_0, r) &= \lim_{n\to\infty} E_1^{(n)}(E_0, r) & \text{(B.37)}\\
&= \lim_{n\to\infty} \min_{\substack{D(Q_1\|P_1)\leq r\\ Q_1\in\mathscr{P}_n(\mathscr{X})}} E_1(E_0, Q_1) & \text{(B.38)}\\
&= \min_{\substack{D(Q_1\|P_1)\leq r\\ Q_1\in\mathscr{P}(\mathscr{X})}} E_1(E_0, Q_1). & \text{(B.39)}
\end{aligned}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Observe that the above limit exists. This is because for $r \in [0, r_c - \varepsilon]$, Q_1 cannot lie on the boundaries of probability simplex, $Q_1(x) > 0 \; \forall x \in \mathscr{X}$, and the function $E_1^{(n)}(E_0, r)$ is continuous. Indeed, for $r \in [0, r_c - \varepsilon]$, the condition $Q_1 \in \mathscr{P}_{\delta_n}(\mathscr{X})$ in the definition of $E_1^{(n)}(E_0, r)$ is inactive. Next, we use the following lemma to find a lower bound on type-II error exponent.

Lemma B.1 For every $n \in \mathbb{Z}^+$, the exponent functions $E_1^{(n)}(E_0, r)$, and $E_1(E_0, r)$ are continuous in $r \in [0, r_c)$.

The proof of Lemma B.1 relies on Berge's maximum theorem [1] and can be found in Appendix E. Taking the limit of $E_1^{(n)}(\phi_\beta)$ when n tends to infinity, we have

$$\begin{aligned} E_1(\phi_\beta) &= \liminf_{n\to\infty} E_1^{(n)}(\phi_\beta) \\ &\geq \liminf_{n\to\infty} \min\left\{ \min_{0\leq r\leq r_c-\varepsilon} E_1^{(n)}(E_0, r) + \alpha r, \alpha(r_c - \varepsilon)\right\} && \text{(B.40)} \\ &= \min\left\{ \min_{0\leq r\leq r_c-\varepsilon} E_1(E_0, r) + \alpha r, \alpha(r_c - \varepsilon)\right\}. && \text{(B.41)} \end{aligned}$$

In order to show (B.41), by Lemma B.1, $E_1^{(n)}(E_0, r)$ and $E_1(E_0, r)$ are continuous on the compact interval $[0, r_c - \varepsilon]$. Also, since $E_1^{(n)}(E_0, r) \geq E_1^{(n+1)}(E_0, r)$ for every r in the domain, by Dini's theorem, the convergence of $E_1^{(n)}(E_0, r)$ is uniform [34]. Finally, since the convergence is uniform, we can interchange the minimum and limit to get (B.41).

When $r = 0$, we have $Q_1 = P_1$ and $\gamma(E_0, P_1) = \beta E_1^*(E_0) - E_0$, resulting $E_1(E_0, r = 0) = E_1^*(E_0)$, where $E_1^*(E_0)$ is the optimal error exponent tradeoff in (2.6). Moreover, for $r_1 \leq r_2$, we have $\mathscr{B}(P_1, r_1) \subseteq \mathscr{B}(P_1, r_2)$, where $\mathscr{B}(P, r) = \{Q \in \mathscr{P}(\mathscr{X}) : D(Q\|P) \leq r\}$. Hence, $E_1(E_0, r)$ is a non-increasing function of r. Therefore, a sufficient condition to have $E_1(\phi_\beta) = E_1^*(E_0)$ is

$$\min\left\{ \min_{0\leq r\leq r_c-\varepsilon} E_1(E_0, r) + \alpha r, \alpha(r_c - \varepsilon)\right\} = E_1(E_0, r = 0). \tag{B.42}$$

Equivalently, letting ε tend to zero, (B.42) can be written as the following two sufficient conditions:

$$E_1(E_0, r) + \alpha r \geq E_1(E_0, r = 0) \; \forall r : \; 0 \leq r < r_c, \tag{B.43}$$

$$\alpha > \frac{E_1(E_0, r = 0)}{r_c}. \tag{B.44}$$

Hence, letting

$$\alpha_\beta^* = \max\left\{ \sup_{0\leq r<r_c} \frac{E_1(E_0, r = 0) - E_1(E_0, r)}{r}, \frac{E_1^*(E_0)}{r_c}\right\} \tag{B.45}$$

for any $\alpha > \alpha_\beta^*$, we have $E_1(\phi_\beta) = E_1^*(E_0)$. Finally, since $E_1(E_0, r)$ is non-increasing and continuous on $r \in [0, r_c)$, and $-\frac{\partial E_1(E_0,r)}{\partial r}\big|_{r=0} < \infty$ as it will be proved in Theorem 3.3, the supremum in (B.45) is

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

finite and

$$0 < \alpha^*_\beta < \infty. \tag{B.46}$$

Next, by using the refinements of large deviation techniques, we show that the proposed classifier achieves the same prefactor as the type-II error probability of the likelihood ratio test, establishing the optimality of the proposed classifier up to a constant. Using the method of types, the type-II probability of error can be written as

$$\varepsilon_1(\phi_\beta) = \sum_{\boldsymbol{x},\boldsymbol{z}:\phi_\beta(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})=0} P_1(\boldsymbol{z})P_1(\boldsymbol{x}) \tag{B.47}$$

$$= \sum_{Q_1\in\mathscr{P}_k(\mathscr{X})\cap\mathscr{P}_{\delta_n}(\mathscr{X})} P_1(\mathscr{T}^k_{Q_1}) \sum_{\substack{Q\in\mathscr{P}_n(\mathscr{X}),\\ \phi_\beta(Q,Q_1)=0}} P_1(\mathscr{T}^n_Q) \tag{B.48}$$

$$\asymp \sum_{Q_1\in\mathscr{P}_k(\mathscr{X})\cap\mathscr{P}_{\delta_n}(\mathscr{X})} ck^{\frac{-|\mathscr{X}|+1}{2}} e^{-kD(Q_1\|P_1)} \sum_{\substack{Q\in\mathscr{P}_n(\mathscr{X}),\\ \phi_\beta(Q,Q_1)=0}} P_1(\mathscr{T}^n_Q), \tag{B.49}$$

where in the last step we used that [12]

$$P_1(\mathscr{T}^k_{Q_1}) \asymp k^{\frac{-|\mathscr{X}|+1}{2}} e^{-kD(Q_1\|P_1)}. \tag{B.50}$$

We need to show that conditioned on the training sequence having a type Q_1, the conditional type-II error probability is

$$\sum_{Q:\phi_\beta(Q,Q_1)=0} P_1(\mathscr{T}^n_Q) = \frac{1}{\sqrt{n}} e^{-nE_1(E_0,Q_1)}\big(c+o(1)\big), \tag{B.51}$$

when $E_1(E_0, Q_1)$ is positive, i.e. the error decay has a prefactor of $\frac{1}{\sqrt{n}}$. We first show that for any classifier with $0 \le \beta \le 1$, we can upper bound the type-II error probability by the classifier type-II error probability with $\beta = 1$.

Lemma B.2 For every fixed Q_1 and for every $\beta_1 \le \beta_2$, we have $\mathscr{A}_0(Q_1, \beta_1) \subseteq \mathscr{A}_0(Q_1, \beta_2)$.

Proof. Let $Q \in \mathscr{A}_0(Q_1, \beta_1)$, we need to show that for every such Q, we have $Q \in \mathscr{A}_0(Q_1, \beta_2)$. We have that

$$D(Q\|P_0) - \beta_1 D(Q\|Q_1) \le D(Q_{\mu^*}\|P_0) - \beta_1 D(Q_{\mu^*}\|Q_1), \tag{B.52}$$

where Q_{μ^*} is the solution to optimization in $\gamma(E_0, Q_1)$ in (B.1) and is independent of β. Furthermore,

$$D(Q\|P_0) - \beta_2 D(Q\|Q_1) \le D(Q_{\mu^*}\|P_0) - \beta_2 D(Q_{\mu^*}\|Q_1) + (\beta_2-\beta_1)\Big(D(Q_{\mu^*}\|Q_1) - D(Q\|Q_1)\Big). \tag{B.53}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Using the KKT conditions, it can be shown that the projection of Q_1 on the set $\mathscr{A}(Q_1,\beta)$ also equals to Q_{μ^*}, i.e.

$$Q_{\mu^*} = \mathop{\arg\min}_{D(Q\|P_0)-\beta_1 D(Q\|Q_1)\le\gamma(E_0,Q_1)} D(Q\|Q_1). \tag{B.54}$$

Hence, for every $Q \in \mathscr{A}_0(Q_1,\beta_1)$, we have

$$D(Q_{\mu^*}\|Q_1) \le D(Q\|Q_1), \tag{B.55}$$

and by (B.53)

$$D(Q\|P_0) - \beta_2 D(Q\|Q_1) \le D(Q_{\mu^*}\|P_0) - \beta_2 D(Q_{\mu^*}\|Q_1), \tag{B.56}$$

i.e. $Q \in \mathscr{A}_0(Q_1,\beta_2)$ which concludes the proof. □

Therefore, we can upper bound the error probability by setting $\beta = 1$. Since for every fixed Q_1, the test is the Neyman–Pearson test using the mismatched distribution Q_1, the decision region boundary is characterised by a hyperplane and by [8] the conditional type-II error probability prefactor equals to $\frac{1}{\sqrt{n}}$ for every Q_1. Hence, substituting (B.51) into (B.49), we have

$$\varepsilon_1(\phi_\beta) \le \sum_{Q_1\in\mathscr{P}_k(\mathscr{X})\cap\mathscr{P}_{\delta_n}(\mathscr{X})} \frac{k^{\frac{-|\mathscr{X}|+1}{2}}}{\sqrt{n}} e^{-kD(Q_1\|P_1)} e^{-nE_1(E_0,Q_1)}\big(c+o(1)\big). \tag{B.57}$$

Observe that by choosing some small enough $\delta_n > 0$, and constraining $\hat{T}_z \in \mathscr{P}_{\delta_n}(\mathscr{X})$, the approximation error caused by replacing $\hat{T}_{z'}$ by $\hat{T}_z$ can be absorbed in the $c + o(1)$ term in (B.57), and we get

$$\varepsilon_1(\phi_\beta) \asymp n^{-\frac{|\mathscr{X}|}{2}} \sum_{Q_1\in\mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)}, \tag{B.58}$$

$$\psi(Q_1) = \frac{1}{\alpha}E_1(E_0,Q_1) + D(Q_1\|P_1), \tag{B.59}$$

where we used the fact that $\alpha = \frac{k}{n}$ is fixed and independent of n, k.

Next, we approximate the summation by an integral using Theorem A.1. As shown earlier in the proof,

$$P_1 = \mathop{\arg\min}_{Q_1:D(Q_1\|P_1)<r_c} \psi(Q_1), \tag{B.60}$$

for $\alpha^*_\beta < \alpha$. We show that the terms in the summation (B.58) where Q_1 is in the vicinity of the P_1 dominate the summation. Let $\mathscr{D} = \mathscr{P}(\mathscr{X})$ and $\mathscr{D}\prime = \mathscr{B}(P_1,\varepsilon)$, where $\varepsilon = \frac{s\log k}{k}$, $s > 0$. We can split

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

the summation (B.58) as

$$\sum_{Q_1 \in \mathscr{D} \cap \mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)} = \sum_{Q_1 \in \mathscr{D}' \cap \mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)} + \sum_{Q_1 \in (\mathscr{D}')^c \cap \mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)}. \tag{B.61}$$

By Theorem 3.3, $\psi(Q_1)$ achieves its minimum at $Q_1 = P_1$ for $\alpha > \alpha^*_\beta$, i.e. $\psi(Q_1) > \psi(P_1)$ for $Q_1 \neq P_1$. Hence, by the mean value theorem and using a Taylor series expansion of $\psi(Q)$ around $Q_1 = P_1$ for $Q_1 \in (\mathscr{D}')^c$, we get

$$\psi(Q_1) = \psi(P_1) + \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \tilde{\boldsymbol{H}}(Q_1)\boldsymbol{\theta}_{P_1}, \tag{B.62}$$

where $\boldsymbol{\theta}_{P_1}$ is the difference vector $Q_1 - P_1$ defined in (C.15), and $\tilde{\boldsymbol{H}}(Q_1)$ is a positive definite matrix for every $Q_1 \in (\mathscr{D}')^c$. Also, we have used the result from Theorem 3.3 that the first-order term of the expansion is zero. Furthermore, note that by the condition $Q_1 \in (\mathscr{D}')^c$, we have $\|\boldsymbol{\theta}_{P_1}\|_\infty^2 = \Omega(\varepsilon)$. Hence, for large enough n, we have

$$\sum_{Q_1 \in (\mathscr{D}')^c \cap \mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)} \leq (k+1)^{\left|\mathscr{X}\right|} e^{-k\left(\frac{1}{\alpha}E_1(E_0,P_1) + \frac{1}{2}\Lambda\frac{s\log k}{k}\right)}, \tag{B.63}$$

where $\Lambda = \min_{Q_1 \in (\mathscr{D}')^c} \Lambda_{\min}\left(\tilde{\boldsymbol{H}}(Q_1)\right)$, and $\Lambda_{\min}(.)$ is the smallest eigenvalue of a matrix. Note that Λ is strictly positive since $\psi(Q_1) > \psi(P_1)$ for $Q_1 \neq P_1$. Finally, by setting $s \geq \frac{2}{\Lambda}(\left|\mathscr{X}\right| + \frac{3}{2})$, we have

$$\sum_{Q_1 \in (\mathscr{D}')^c \cap \mathscr{P}_k(\mathscr{X})} e^{-k\psi(Q_1)} \leq n^{-\frac{1}{2}} e^{-nE_1(E_0,P_1)} O\left(n^{-1}\right). \tag{B.64}$$

It remains to bound the first term in (B.61). By the envelope theorem [27] as shown in Lemma C.2, we can conclude that $\psi : \mathbb{R}^{\left|\mathscr{X}\right|-1} \to \mathbb{R}$ is continuously differentiable and hence Lipschitz on $\mathscr{D}'$. Furthermore, $\mathscr{D}'$ is a C^∞ manifold, and hence, $\mathscr{D}'$ has a minimally smooth boundary. Therefore, by Theorem A.1, we have

$$\varepsilon_1(\phi_\beta) \asymp n^{\frac{\left|\mathscr{X}\right|-2}{2}} \int_{\mathscr{D}'} e^{-k\psi(\boldsymbol{w})}\, d\boldsymbol{w}. \tag{B.65}$$

By using a Taylor series expansion of $\psi(\boldsymbol{w})$ around $\boldsymbol{w} = P_1$, and Lemma C.2, we have

$$\varepsilon_1(\phi_\beta) \asymp n^{\frac{\left|\mathscr{X}\right|-2}{2}} \int_{\mathscr{D}'} e^{-n\left(E_1(E_0,P_1) + \frac{1}{2}\langle(\boldsymbol{w}-P_1),(\boldsymbol{H}_2+\alpha\boldsymbol{J})(\boldsymbol{w}-P_1)\rangle + O\left(\varepsilon^{\frac{3}{2}}\right)\right)} d(\boldsymbol{w} - P_1). \tag{B.66}$$

Note that $\boldsymbol{w}$ is bounded to the $\left|\mathscr{X}\right| - 1$ dimensional probability simplex as $\boldsymbol{w} \in \mathscr{D}'$. Hence, for every $\boldsymbol{w} - P_1$, there exists a $\left|\mathscr{X}\right| \times \left|\mathscr{X}\right|$ rotation matrix $\boldsymbol{R}$ such that for every $\boldsymbol{w} \in \mathscr{P}(\mathscr{X})$, we have $\boldsymbol{e}^T_{\left|\mathscr{X}\right|}\boldsymbol{R}(\boldsymbol{w} -$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

$P_1) = 0$, where $\boldsymbol{e}_{|\mathscr{X}|}$ is the unit vector with 1 in the $|\mathscr{X}|$'th coordinate and 0 otherwise. Letting $\boldsymbol{w}\prime = \boldsymbol{R}(\boldsymbol{w} - P_1)$, we have

$$\varepsilon_1(\phi_\beta) \asymp n^{\frac{|\mathscr{X}|-2}{2}} \int_{\mathscr{D}\cap\mathbb{R}^{|\mathscr{X}|-1}} e^{-n\left(E_1(E_0,P_1)+\frac{1}{2}\left\langle \boldsymbol{R}^T\boldsymbol{w}', (\boldsymbol{H}_2+\alpha\boldsymbol{J})\boldsymbol{R}^T\boldsymbol{w}'\right\rangle + O\left(\varepsilon^{\frac{3}{2}}\right)\right)} d\boldsymbol{R}^T\boldsymbol{w}', \tag{B.67}$$

where we have used $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$. Furthermore, from the proof of Theorem 3.3, $\boldsymbol{H}_2 + \alpha\boldsymbol{J}$ is strictly positive definite and hence full rank for $\alpha^*_\beta < \alpha$. Using a change of variables $\boldsymbol{y} = \sqrt{n}\boldsymbol{w}\prime$, we have $d\boldsymbol{y} = n^{-\frac{|\mathscr{X}|-1}{2}} d\boldsymbol{w}\prime$, and

$$\varepsilon_1(\phi_\beta) \asymp n^{-\frac{1}{2}} e^{-nE_1(E_0,P_1)} \int_{\mathscr{D}} e^{-\frac{1}{2}\left\langle \boldsymbol{R}^T\boldsymbol{y}, (\boldsymbol{H}_2+\alpha\boldsymbol{J})\boldsymbol{R}^T\boldsymbol{y}\right\rangle + O\left(\sqrt{\frac{\log^3 n}{n}}\right)} d\boldsymbol{y}, \tag{B.68}$$

where the integral is a Gaussian integral and equals to $c\prime + o(1)$ for a constant $c\prime$ and this concludes the proof.

C. Proof of Thereom 3.3

From (B.45), we have

$$\underline{\alpha}_\beta \triangleq -\frac{\partial E_1(E_0, r)}{\partial r}\bigg|_{r=0} \leq \alpha^*_\beta. \tag{C.1}$$

To find the derivative of $E_1(E_0, r)$ at $r = 0$, we use a Taylor series expansion of $E_1(E_0, Q_1)$ around $Q_1 = P_1$. We use the following lemmas in our proof. First lemma provides the first and second derivative of $\gamma(E_0, Q_1)$.

Lemma C.1 For every $P_0, P_1 \in \mathscr{P}(\mathscr{X})$, the gradient of threshold function $\gamma(E_0, Q_1)$ is

$$\frac{\partial\gamma(E_0, Q_1)}{\partial Q_1(i)} = -\beta\frac{Q_{\mu^*}(i)}{Q_1(i)}. \tag{C.2}$$

Furthermore, the Hessian matrix of $\gamma(E_0, Q_1)$ evaluated at $Q_1 = P_1$ is

$$\boldsymbol{H}_1 = \boldsymbol{H}(Q_1)\Big|_{Q_1=P_1}, \tag{C.3}$$

$$\boldsymbol{H}(Q_1) = \nabla^2_{Q_1}\gamma(E_0, Q)\Big|_{Q_1=P_1} \tag{C.4}$$

$$= \beta\boldsymbol{J}\left(\frac{\mu^*}{1+\mu^*}\boldsymbol{T} + \frac{1}{(1+\mu^*)}(\boldsymbol{Q} + \boldsymbol{W})\right)\boldsymbol{J}, \tag{C.5}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where $\boldsymbol{W} = \boldsymbol{w}\boldsymbol{w}^T, \boldsymbol{Q} = \boldsymbol{q}\boldsymbol{q}^T$ and

$$\boldsymbol{w}^T = \frac{1}{\sqrt{\mathrm{Var}_{Q_{\mu^*}}\left(\log \frac{Q_{\mu^*}}{P_0}\right)}}\Bigg(Q_{\mu^*}(1)\log\left(\frac{Q_{\mu^*}(1)}{P_0(1)}\right) - E_0, \dots, Q_{\mu^*}(|\mathcal{X}|)\log\left(\frac{Q_{\mu^*}(|\mathcal{X}|)}{P_0(|\mathcal{X}|)}\right) - E_0\Bigg), \tag{C.6}$$

$$\boldsymbol{q} = \Big(Q_{\mu^*}(1), \dots, Q_{\mu^*}(|\mathcal{X}|)\Big), \tag{C.7}$$

$$\boldsymbol{T} = \mathrm{diag}\Big(Q_{\mu^*}\Big), \quad \boldsymbol{J} = \mathrm{diag}\left(\frac{1}{Q_1}\right) \tag{C.8}$$

and Q_{μ^*}, μ^* are defined as in (B.1) and (B.2).

Proof. The proof can be found in Appendix F. □

Next using the derivatives of $\gamma(E_0, Q_1)$ and (B.29), we can find the derivatives of the exponent function $E_1(E_0, Q_1)$.

LEMMA C.2 For every $P_0, P_1 \in \mathscr{P}(\mathcal{X})$, the gradient of the exponent function $E_1(E_0, Q_1)$ evaluated at $Q_1 = P_1$ is

$$\left.\frac{\partial E_1(E_0, Q_1)}{\partial Q_1(i)}\right|_{Q_1 = P_1} = 0. \tag{C.9}$$

Furthermore, the Hessian matrix evaluated at $Q_1 = P_1$ is

$$\boldsymbol{H}_2 \triangleq \nabla^2_{Q_1} E_1(E_0, Q_1)\Big|_{Q_1 = P_1} \tag{C.10}$$

$$= \beta\eta^*_\beta \boldsymbol{J}\Big[\boldsymbol{Q} + \eta^*_1 \boldsymbol{V} + (1 - \eta^*_1)\boldsymbol{W} - \boldsymbol{T}\Big]\boldsymbol{J}, \tag{C.11}$$

where η^*_β and η^*_1 are the optimal Lagrange multipliers in (3.15) when $0 < \beta \leq 1$ and $\beta = 1$, respectively, and $\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^T$

$$\boldsymbol{v} = \frac{1}{\sqrt{\mathrm{Var}Q_{\eta^*_\beta}(\Omega)}}\Big(Q_{\eta^*_\beta}(1)\big(\Omega(1) - (E_0 - \beta E_1)\big), \dots, Q_{\eta^*_\beta}(|\mathcal{X}|)\big(\Omega(|\mathcal{X}|) - (E_0 - \beta E_1)\big)\Big), \tag{C.12}$$

$$\Omega(i) = \beta \log \frac{P_1(i)}{P_0(i)} + (1 - \beta)\log \frac{Q_{\eta^*_\beta}(i)}{P_0(i)}, i \in \mathcal{X}. \tag{C.13}$$

Proof. The proof can be found in Appendix G. □

Lemma C.2 shows that the gradient of $E_1(E_0, Q_1)$ respect to Q_1 is zero at $Q_1 = P_1$ which is due to the choice of the threshold function $\gamma(E_0, Q_1)$. In the case of the plugin classifier with the fixed threshold, the gradient is non-zero which results in $\left.\frac{\partial E_1(E_0, r)}{\partial r}\right|_{r=0} = \infty$. Next, by using a Taylor series expansion of $E_1(E_0, Q_1)$ and the relative entropy in (B.39), we can get the following lemma.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Lemma C.3 For any $P_0, P_1 \in \mathscr{P}(\mathscr{X})$, the approximation

$$E_1(E_0, r) = E_1^*(E_0) + \min_{\substack{\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}\boldsymbol{\theta}_{P_1} \le r \\ \mathbf{1}^T\boldsymbol{\theta}_{P_1}=0}} \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{H}_2 \boldsymbol{\theta}_{P_1} + o(r), \tag{C.14}$$

where

$$\boldsymbol{\theta}_{P_1} = \Big(Q_1(1) - P_1(1), \ldots, Q_1(\left|\mathscr{X}\right|) - P_1(\left|\mathscr{X}\right|)\Big)^T \tag{C.15}$$

holds.

Proof. The proof can be found in Appendix H. □

We can further simplify the convex optimization problem in (C.14), using the following lemma.

Lemma C.4 For $\boldsymbol{J}, \boldsymbol{H}_2$ defined in (3.9), (C.10), we have

$$\min_{\substack{\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}\boldsymbol{\theta}_{P_1} \le r \\ \mathbf{1}^T\boldsymbol{\theta}_{P_1}=0}} \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{H}_2 \boldsymbol{\theta}_{P_1} = \min_{\frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{\psi}_{P_1} \le r} \frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1}, \tag{C.16}$$

where

$$\boldsymbol{H} = \beta\eta_\beta^* \sqrt{\boldsymbol{J}}\Big[\boldsymbol{Q} + \eta_1^*\boldsymbol{V} + (1-\eta_1^*)\boldsymbol{W} - \boldsymbol{T}\Big]\sqrt{\boldsymbol{J}}. \tag{C.17}$$

Proof. The proof can be found in Appendix I. □

Finally, by Lemmas C.3 and C.4, we only need to solve the optimization in (C.16). Assuming $\boldsymbol{H}$ has negative eigenvalues which is the case for $\left|\mathscr{X}\right| \geq 4$ by Weyl's inequality [19], it can be shown that at the optimal solution $\boldsymbol{\psi}_{P_1}^*$, the inequality constraint of optimization problem (C.16) is satisfied with equality, hence

$$\min_{\frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{\psi}_{P_1} \le r} \frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1} = r \min_{\frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{\psi}_{P_1} \le r} \frac{1}{2} \frac{\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1}}{\boldsymbol{\psi}_{P_1}^T \boldsymbol{\psi}_{P_1}} \tag{C.18}$$

$$= \Lambda_{\min}(\boldsymbol{H}) r, \tag{C.19}$$

where $\Lambda_{\min}(\boldsymbol{H})$ is the smallest eigenvalue of $\boldsymbol{H}$. Taking the derivative of (C.14) and setting $r = 0$, we obtain

$$\underline{\alpha}_\beta = -\Lambda_{\min}(\boldsymbol{H}). \tag{C.20}$$

Further upper bounding $\Lambda_{\min}(\boldsymbol{H})$ and further using Weyl's inequality [19], we get

$$-\Lambda_{\min}(\boldsymbol{H}) \geq \beta\eta_\beta^* \left[\frac{Q_{\eta_\beta^*}}{P_1}\right]_{(3)}, \tag{C.21}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where we have used the fact that $\boldsymbol{Q} + \eta_1^* \boldsymbol{V} + (1 - \eta_1^*)\boldsymbol{W}$ can only have three non-zero eigenvalues as $\boldsymbol{Q}, \boldsymbol{V}, \boldsymbol{W}$ are rank one matrices.

D. Proof of Theorem 3.5

By Lemma B.2, we can see that for every $\beta_1 \le \beta_2$, we have $\mathscr{A}_0(Q_1, \beta_1) \subseteq \mathscr{A}_0(Q_1, \beta_2)$, and hence, $E_1(E_0, Q_1)$ is non-increasing in β. Therefore, in order to find an upper bound for α_β^*, we can find an upper bound to $\alpha^* = \alpha_{\beta=1}^*$, which would be an upper bound to any test with $\beta < 1$. Hence, we set $\beta = 1$ in the rest of the proof.

In order to upper bound α^* in (B.45), we first find a lower bound to $E_1(E_0, r)$. Rewriting $E_1(E_0, Q_1)$ in its dual form, we have [3]

$$E_1(E_0, Q_1) = \max_{\nu \ge 0} -\nu\gamma(E_0, Q_1) - \log \sum_{a \in \mathscr{X}} P_0^{\nu}(a) P_1(a) Q_1^{-\nu}(a). \tag{D.1}$$

Setting $\nu = \lambda^*$, where λ^* is the optimal Lagrange multiplier in (2.11) when P_1 is known in the test. We get

$$E_1(E_0, Q_1) \ge -\lambda^* \gamma(E_0, Q_1) - \log \sum_{a \in \mathscr{X}} P_0^{\lambda^*}(a) P_1(a) Q_1^{-\lambda^*}(a). \tag{D.2}$$

By the Taylor remainder theorem, for every $Q_1 \in \mathscr{B}(P_1, r)$ where $r < r_c$, we have

$$E_1(E_0, Q_1) \ge E_1(E_0, P_1) + \frac{1}{2} \boldsymbol{\theta}_{P_1}^T \boldsymbol{\Sigma}(\tilde{Q}_1) \boldsymbol{\theta}_{P_1}, \tag{D.3}$$

where

$$\boldsymbol{\Sigma}(Q_1) = -\lambda^* \boldsymbol{H}(Q_1) - \lambda^*(1 + \lambda^*) \operatorname{diag}\left(\frac{\hat{Q}_{\lambda^*}}{Q_1^2}\right) + (\lambda^*)^2 \boldsymbol{u}^T \boldsymbol{u}, \tag{D.4}$$

$$\boldsymbol{H}(Q_1) = \nabla_{Q_1}^2 \gamma(E_0, Q), \tag{D.5}$$

$$\boldsymbol{u} = \left(\frac{\hat{Q}_{\lambda^*}(1)}{Q_1(1)}, \ldots, \frac{\hat{Q}_{\lambda^*}(k)}{Q_1(k)}\right)^T, \tag{D.6}$$

$$\hat{Q}_{\lambda^*}(x) = \frac{P_0^{\lambda^*}(x) P_1(x) Q_1^{-\lambda^*}(x)}{\sum_{a \in \mathscr{X}} P_0^{\lambda^*}(a) P_1(a) Q_1^{-\lambda^*}(a)}, \tag{D.7}$$

evaluated at some $\tilde{Q}_1 \in \mathscr{B}(P_1, r)$, and $\boldsymbol{\theta}_{P_1}$ is difference vector $Q_1 - P_1$ defined in (C.15). $\boldsymbol{H}(Q_1)$ is the Hessian matrix of the threshold function in (C.5). We have used the fact that the gradient of RHS of (D.2) evaluated at $Q_1 = P_1$ is equal to zero, hence we only need to control the second-order term. Letting

$$\varLambda(\boldsymbol{\Sigma}, r) = \min_{\tilde{Q}_1 \in \mathscr{B}(P_1, r)} \varLambda_{\min}\big(\boldsymbol{\Sigma}(\tilde{Q}_1)\big), \tag{D.8}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where $\Lambda_{\min}\big(\boldsymbol{\Sigma}(\tilde{Q}_1)\big)$ is the smallest eigenvalue of $\boldsymbol{\Sigma}(\tilde{Q}_1)$. We can further lower bound $E_1(E_0, Q_1)$ by

$$E_1(E_0, Q_1) \geq E_1(E_0, P_1) + \frac{\Lambda(\boldsymbol{\Sigma}, r)}{2}\boldsymbol{\theta}_{P_1}^T\boldsymbol{\theta}_{P_1}. \tag{D.9}$$

Substituting (D.9) in (B.39), we get

$$E_1(E_0, r) \geq \min_{\substack{D(Q_1\|P_1)\leq r\\ Q_1\in\mathscr{P}(\mathscr{X})}} E_1(E_0, P_1) + \frac{\Lambda(\boldsymbol{\Sigma}, r)}{2}\boldsymbol{\theta}_{P_1}^T\boldsymbol{\theta}_{P_1}. \tag{D.10}$$

Similarly to $E_1(E_0, Q_1)$, by using a Taylor series expansion of $D(Q_1\|P_1)$ around $Q_1 = P_1$ and using the remainder theorem, for every $Q_1 \in \mathscr{B}(P_1, r)$, we have

$$D(Q_1\|P_1) \geq \frac{\Lambda(\boldsymbol{J}, r)}{2}\boldsymbol{\theta}_{P_1}^T\boldsymbol{\theta}_{P_1}, \tag{D.11}$$

$$\boldsymbol{J}(\tilde{Q}_1) = \mathrm{diag}\left(\frac{1}{\tilde{Q}_1(1)}, \dots, \frac{1}{\tilde{Q}_1(|\mathscr{X}|)}\right). \tag{D.12}$$

Therefore, by (D.10) and (D.11), we get

$$E_1(E_0, r) \geq E_1(E_0, P_1) + \frac{\Lambda(\boldsymbol{\Sigma}, r)}{\Lambda(\boldsymbol{J}, r)}r. \tag{D.13}$$

By Weyl's inequality [19], we have

$$\Lambda(\boldsymbol{\Sigma}, r) \geq \min_{\tilde{Q}_1\in\mathscr{B}(P_1,r)} -\lambda^*\left(\frac{\mu^*}{1+\mu^*}\left[\frac{Q_{\mu^*}}{\tilde{Q}_1^2}\right]_{(1)} + \frac{1}{(1+\mu^*)}\big(\|\boldsymbol{w}_{\tilde{Q}_1}\|_2^2 + \|\boldsymbol{q}_{\tilde{Q}_1}\|_2^2\big)\right) - \lambda^*(1+\lambda^*)\left[\frac{Q_{\lambda^*}}{\tilde{Q}_1^2}\right]_{(1)} \tag{D.14}$$

$$\geq -\lambda^*(4+\lambda^*)\left(\frac{1}{Q_1^{\min}(r)}\right)^2, \tag{D.15}$$

where

$$Q_1^{\min}(r) = \min_{\substack{\tilde{Q}_1\in\mathscr{B}(P_1,r)\\ a\in\mathscr{X}}} \tilde{Q}_1(a), \tag{D.16}$$

and we used $Q_{\lambda^*}(a), Q_{\mu^*}(a) \leq 1$ for all $a \in \mathscr{X}$ and $\mu^* \geq 0$. Finally, by $\Lambda(\boldsymbol{J}, r) \geq 1$, we get

$$E_1(E_0, r) \geq E_1(E_0, P_1) - \lambda^*(4+\lambda^*)\left(\frac{1}{Q_1^{\min}(r)}\right)^2 r. \tag{D.17}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

By Pinsker's inequality [11], we have $\|\boldsymbol{\theta}_{P_1}\|_1 \leq \sqrt{2r}$, hence $Q_1^{\min} \geq P_1^{\min} - \frac{\sqrt{2r}}{2}$, and

$$E_1(E_0, r) \geq E_1(E_0, P_1) - \lambda^*(4+\lambda^*)\left(\frac{\sqrt{r}}{P_1^{\min} - \sqrt{r}}\right)^2. \tag{D.18}$$

Since $E_1(E_0, r) \geq 0$ for all $r \geq 0$, we can improve the lower bound to

$$E_1(E_0, r) \geq \left\{E_1(E_0, P_1) - \lambda^*(4+\lambda^*)\left(\frac{\sqrt{r}}{P_1^{\min} - \sqrt{r}}\right)^2\right\}\mathbb{1}\{r \leq \bar{r}\}, \tag{D.19}$$

where

$$\bar{r} = \left(\frac{\kappa P_1^{\min}}{1+\kappa}\right)^2. \tag{D.20}$$

Next by (B.45) and (D.19), we have

$$\alpha_\beta^* \leq \max\left\{\max_{0\leq r\leq r_c} \lambda^*(4+\lambda^*)\left(\frac{\mathbb{1}\{r \leq \bar{r}\}}{P_1^{\min} - \sqrt{r}}\right)^2, \frac{E_1(E_0, P_1)}{r_c}\right\} \tag{D.21}$$

$$= \frac{\lambda^*(4+\lambda^*)}{(P_1^{\min} - \sqrt{\bar{r}})^2} \tag{D.22}$$

$$= \frac{\lambda^*(4+\lambda^*)(1+\kappa)}{(P_1^{\min})^2}, \tag{D.23}$$

where we have dropped the second term in the maximization as it is derived by letting $E_1(E_0, r)$ to be a straight line from $E_1(E_0, r=0)$ at $r = 0$ to 0 at $r = r_c$ and the first term in the maximization is derived by the lower bound to $E_1(E_0, r)$, and $\bar{r} < r_c$, hence the first term dominates. This concludes the proof.

E. Proof of Lemma B.1

We need to prove the the continuity of $E_1(E_0, r)$ on $r \in [0, r_c)$. To prove this, we will use the Berge's Maximum Theorem stated below [1].

Proposition E.1 Berge's Maximum TheoremLet $X \subseteq \mathbb{R}^n$ and $\Theta \subseteq \mathbb{R}^m$. Let $F : X \times \Theta \to \mathbb{R}$ be a continuous function, and let $\mathscr{G} : \Theta \to X$ be a compact valued and continuous correspondence. Then, the maximum value function

$$V(\Theta) = \max_{X \in \mathscr{G}(\Theta)} F(X, \Theta) \tag{E.1}$$

is well-defined and continuous, and the optimal policy correspondence

$$X^*(\Theta) = \{X \in \mathscr{G}(\Theta) \big| F(X, \Theta) = V(\Theta)\} \tag{E.2}$$

is non-empty, compact valued and upper hemicontinuous.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

We first prove that correspondence $\mathscr{G}_1(Q_1) = \{Q \in \mathscr{P}(\mathscr{X}) : D(Q\|P_0) - D(Q\|Q_1) \geq \gamma(E_0, Q_1)\}$ is a continuous correspondence for $Q_1 \in \mathscr{P}_\delta(\mathscr{X})$. Then, by Berge's theorem, $E_1(E_0, Q_1)$ is continuous on Q_1 over $\mathscr{P}_\delta(\mathscr{X})$. Next, by showing the continuity of the correspondence $\mathscr{G}_2(r) = \{Q_1 \in \mathscr{P}_\delta(\mathscr{X}) : D(Q_1\|P_1) \leq r\}$, we conclude the continuity of $E_1(E_0, r)$ in r by Berge's maximum theorem. Therefore, we only need to prove that $\mathscr{G}_1, \mathscr{G}_2$ are compact and continuous correspondences. Note that Q and Q_1 are subsets of $\mathbb{R}^{|\mathscr{X}|}$ and also $\mathscr{G}_1(Q_1), \mathscr{G}_2(Q)$ are both closed and bounded, hence by the Heine–Borel theorem [34], both $\mathscr{G}_1, \mathscr{G}_2$ are compact. To prove the continuity of a correspondence, we need to prove the upper and lower hemicontinuity of the correspondence.

Definition E.2 The compact valued correspondence $\mathscr{G}$ is upper hemicontinuous at $\theta \in \Theta$ if $\mathscr{G}(\theta)$ is non-empty and if, for every sequence $\{\theta^{(j)}\}$ with $\theta^{(j)} \to \theta$ and every sequence $\{X^{(j)}\}$ with $X^{(j)} \in \mathscr{G}(\theta^{(j)})$ for all j, there exists a convergent subsequence $\{X^{(j_k)}\}$ such that $X^{(j_k)} \to X \in \mathscr{G}(\theta)$.

To prove upper hemicontinuity of $\mathscr{G}_1(Q_1)$, fix $Q_1 \in \mathscr{P}_\delta(\mathscr{X})$ and assume for every $1 \leq j$, $Q_1^{(j)} \in \mathscr{P}_\delta(\mathscr{X})$ be a sequence converging to Q_1. Moreover, assume for every j, there exists a $Q^{(j)}$ such that $Q^{(j)} \in \mathscr{G}_1(Q_1^{(j)})$ for all j. By the definition of convergence in metric spaces, since $Q_1^{(j)} \to Q_1$, then there exists a closed bounded set $\mathscr{Q}_2 \subseteq \mathscr{P}_\delta$, such that for all large enough j, we have $Q_1^{(j)} \in \mathscr{Q}_2 \subset \mathbb{R}^{|Xc|-1}$. Furthermore, since for every $Q_1^{(j)}, \mathscr{G}_1(Q_1^{(j)})$ is an intersection of a half space created by the classifier and the probability simplex, $Q^{(j)} \in \mathscr{G}_1(Q_1)$ will lie in a closed and bounded subset of $\mathbb{R}^{|\mathscr{X}|-1}$. Therefore, for large enough j, the tuple $(Q^{(j)}, Q_1^{(j)})$ also lies in a closed and bounded subset of $\mathbb{R}^{2|\mathscr{X}|-2}$. Finally, by the Bolzano–Weierstrass theorem [34], each bounded sequence in the Euclidean space $\mathbb{R}^{2|\mathscr{X}|-2}$ has a convergent subsequence $(Q^{(j_k)}, Q_1^{(j_k)})$ with the limit point (Q, Q_1). Finally, since each element of this convergent subsequence satisfies the constraint in $\mathscr{G}_1(Q_1)$, then by showing the continuity of $\gamma(E_0, Q_1)$ in Q_1, we can conclude that the limit point will also satisfy $Q \in \mathscr{G}_1(Q_1)$ and gives the upper hemicontinuity. To prove the continuity of $\gamma(E_0, Q_1)$, observe that $D(P\|Q)$ is continuous in the pair for any $P \in \mathscr{P}(\mathscr{X})$ and $Q \in \mathscr{P}_\delta(\mathscr{X})$ where $\delta > 0$. Hence, by the Berge's theorem, $\gamma(E_0, Q_1)$ is continuous in $Q_1 \in \mathscr{P}_\delta(\mathscr{X})$, and the optimizer $Q_\mu(Q_1)$ is upper hemicontinuous. Also, since by (B.1), $Q_\mu(Q_1)$ is single valued, we get that $Q_\mu(Q_1)$ is in fact continuous in Q_1.

Definition E.3 The correspondence $\mathscr{G}$ is lower hemicontinuous at $\theta \in \Theta$ if $\mathscr{G}(\theta)$ is non-empty and if, for every $X \in \mathscr{G}(\theta)$ and every sequence $\{\theta^{(j)}\}$ such that $\theta^{(j)} \to \theta$, there is a $1 \leq J$ and a sequence $\{X^{(j)}\}$ such that $X^{(j)} \in \mathscr{G}(\theta^{(j)})$ for all $J \leq j$ and $X^{(j)} \to X$.

To prove the lower hemicontinuity of $\mathscr{G}_1(Q_1)$, let $Q_1 \in \mathscr{P}(\mathscr{X})$ and $Q \in \mathscr{G}_1(Q_1)$ be fixed. Also, let $Q_1^{(j)} \in \mathscr{P}(\mathscr{X})$ be a sequence converging to Q_1. Next, we construct the sequence $Q^{(j)}$ such that for every j, $Q^{(j)} \in \mathscr{G}_1(Q_1^{(j)})$ and $Q^{(j)} \to Q$. Note that for every $Q_1^{(j)}$, the correspondence is a half space characterized by the hyperplane

$$\mathscr{H}^{(j)} = \{Q \in \mathscr{P}(\mathscr{X}) : \big(Q - Q_\mu(Q_1^{(j)})\big)^T \boldsymbol{n}^{(j)} = 0\}, \tag{E.3}$$

where $\boldsymbol{n}^{(j)}$ is

$$\boldsymbol{n}^{(j)} = \left(\log \frac{Q_\mu(Q_1^{(j)})(1)}{P_0(1)}, \ldots, \log \frac{Q_\mu(Q_1^{(j)})(|\mathscr{X}|)}{P_0(|\mathscr{X}|)}\right)^T. \tag{E.4}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Also for every $Q^{(j)}$, define the line passing P_0 and intersecting $\mathscr{H}^{(j)}$ as

$$L(Q^{(j)}) = \beta P_0 + (1-\beta)Q_{\dagger}^{(j)}, \beta \in [0,1], \tag{E.5}$$

$$Q_{\dagger}^{(j)} = \{P : P = \beta P_0 + (1-\beta)Q, \beta \in \mathbb{R}\} \cap \mathscr{H}^{j}. \tag{E.6}$$

Moreover, for every Q, let $Q = \beta^* P_0 + (1-\beta^*)Q_{\dagger}$. Finally, we can define the sequence $Q^{(j)}$ as

$$Q^{(j)} = \beta^* P_0 + (1-\beta^*)Q_{\dagger}^{(j)}. \tag{E.7}$$

By our construction, it is clear that for every j, $Q^{(j)} \in \mathscr{G}_1(Q_1^{(j)})$. Moreover, by the continuity of $Q_\mu(Q_1^{(j)})$, we conclude that $\mathscr{H}^{(j)} \to \mathscr{H}$ and consequently $Q_{\dagger}^{(j)} \to Q_{\dagger}$, and $Q^{(j)} \to Q$.

Next, we show the continuity of $\mathscr{G}(r)$. To show the upper hemicontinuity of the correspondence, we can use the same argument as the upper hemicontinuity of $\mathscr{G}_1(Q_1)$. Therefore, it only remains to prove the lower hemicontinuity. Let $0 \le r$ and $Q_1 \in \mathscr{P}_\delta(\mathscr{X})$ be fixed. Also, let $0 \le r^{(j)}$ is a sequence converging to r. Construct $Q_1^{(j)} \in \mathscr{P}_\delta(\mathscr{X})$ as

$$Q_1^{(j)} = \left(\frac{r^{(j)}}{r}\right)Q_1 + \left(1 - \frac{r^{(j)}}{r}\right)P_1. \tag{E.8}$$

It is clear that $Q_1^{(j)} \to Q_1$ as $r^{(j)} \to r$. Also by convexity of KL-divergence, we get

$$D\big(Q_1^{(j)} \| P_1\big) \le \frac{r^{(j)}}{r} D(Q_1 \| P_1) \le r^{(j)} \tag{E.9}$$

where in the last inequality, we used the fact that $Q_1 \in \mathscr{G}(r)$. Therefore, $Q_1^{(j)} \in \mathscr{G}\big(r^{(j)}\big)$, gives the lower hemicontinuity of $\mathscr{G}(r)$ and concludes the proof.

F. Proof of Lemma C.1

By the envelope theorem [27], we can find the partial derivative simply by taking the derivative of Lagrangian and evaluating at its optimal solution. Writing the Lagrangian, we have

$$L(Q, Q_1, \mu, \nu) = D(Q\|Q_1) + \mu\big(D(Q\|P_0) - E_0\big) + \nu\left(\sum_{x\in\mathscr{X}} Q(x) - 1\right), \tag{F.1}$$

hence

$$\frac{\partial \gamma}{\partial Q_1(j)} = \beta \frac{\partial L}{\partial Q_1(j)}\bigg|_{Q=Q_{\mu^*}} \tag{F.2}$$

$$= -\beta \frac{Q_{\mu^*}(j)}{Q_1(j)}. \tag{F.3}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Taking the second-order partial derivative, we get

$$\frac{\partial^2 \gamma}{\partial Q_1(i) Q_1(j)} = \beta\left(- \frac{1}{Q_1(j)} \frac{\partial Q_{\mu^*}(j)}{\partial Q_1(i)} + \frac{Q_{\mu^*}(i)}{Q_1^2(i)} \mathbb{1}\{i = j\}\right). \tag{F.4}$$

Therefore, we need to find the sensitivity of Q_{μ^*}, the optimal solution of optimization (3.1), to local changes in Q_1. Reference [7] presents a general approach to find the partial derivative of primal and dual variables solution with respect to any parameter of the optimization problem. For ease of reference, we state this result. The result in [7] is more general; however, we only state the version we need in our proof.

Proposition F.1 Consider the following primal non-linear programming problem:

$$\min_{\substack{\boldsymbol{x}:\, g(\boldsymbol{x},\boldsymbol{a}) \le 0 \\ h(\boldsymbol{x},\boldsymbol{a}) = 0}} z = f(\boldsymbol{x}, \boldsymbol{a}), \tag{F.5}$$

where $f, g, h : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, and $f, h, g \in C^2$. Let μ, ν be the Lagrange multipliers corresponding to inequality and equality constraint. Furthermore, assume at the optimal solution, the constraints are active and $\mu^* \neq 0$, then

$$\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{a}}, \frac{\partial \nu}{\partial \boldsymbol{a}}, \frac{\partial \mu}{\partial \boldsymbol{a}}, \frac{\partial z}{\partial a} \right)^T = \boldsymbol{U}^{-1} \boldsymbol{S}, \tag{F.6}$$

where

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{F}_{xx} & \boldsymbol{H}_x & \boldsymbol{G}_x & \boldsymbol{0} \\ \boldsymbol{H}_x^T & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{G}_x^T & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{F}_x^T & \boldsymbol{0} & \boldsymbol{0} & -1 \end{pmatrix}, \quad \boldsymbol{S} = -\begin{pmatrix} \boldsymbol{F}_{xa} \\ \boldsymbol{H}_a^T \\ \boldsymbol{G}_a^T \\ \boldsymbol{F}_a^T \end{pmatrix} \tag{F.7}$$

and $\boldsymbol{0}$ is the matrix with entries equal to zero with the corresponding dimensions and the submatrices are defined as

$$\begin{aligned} \boldsymbol{F}_{xx} &= \nabla_{xx} f(x^*, a) + \mu^* \nabla_{xx} g(x^*, a) + \nu^* \nabla_{xx} h(x^*, a), \\ \boldsymbol{F}_{xa} &= \nabla_{xa} f(x^*, a) + \mu^* \nabla_{xa} g(x^*, a) + \nu^* \nabla_{xa} h(x^*, a), \end{aligned} \tag{F.8}$$

where (x^*, μ^*, ν^*) is the optimal primal and dual variable solutions. Moreover, $\boldsymbol{F}_x, \boldsymbol{F}_a, \boldsymbol{G}_x, \boldsymbol{G}_a, \boldsymbol{H}_x, \boldsymbol{H}_a$ are gradient of f, g, h respect to x, a evaluated at x^*, respectively.

Using this proposition, we can find $\frac{\partial Q_{\mu^*}}{\partial Q_1}$ in equation (F.4). Letting $x = Q$, and $a = Q_1$, we have

$$\boldsymbol{F}_{QQ} = (1 + \mu^*) \mathrm{diag}\left(\frac{1}{Q_{\mu^*}(1)}, \ldots, \frac{1}{Q_{\mu^*}(|\mathscr{X}|)} \right), \tag{F.9}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

$$\boldsymbol{F}_{QQ_1} = -\mathrm{diag}\Big(\frac{1}{Q_1(1)}, \ldots, \frac{1}{Q_1(|\mathscr{X}|)}\Big), \tag{F.10}$$

$$\boldsymbol{F}_{Q} = \left(1 + \log \frac{Q_{\mu^*}(1)}{Q_1(1)}, \ldots, 1 + \log \frac{Q_{\mu^*}(|\mathscr{X}|)}{Q_1(|\mathscr{X}|)}\right)^T, \tag{F.11}$$

$$\boldsymbol{F}_{Q_1} = \left(-\frac{Q_{\mu^*}(1)}{Q_1(1)}, \ldots, -\frac{Q_{\mu^*}(|\mathscr{X}|)}{Q_1(|\mathscr{X}|)}\right)^T, \tag{F.12}$$

$$\boldsymbol{G}_{Q} = \left(1 + \log \frac{Q_{\mu^*}(1)}{P_0(1)}, \ldots, 1 + \log \frac{Q_{\mu^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)}\right)^T, \tag{F.13}$$

$$\boldsymbol{G}_{Q_1} = (0, \ldots, 0)^T, \tag{F.14}$$

$$\boldsymbol{H}_{Q} = (1, \ldots, 1)^T, \tag{F.15}$$

$$\boldsymbol{H}_{Q_1} = (0, \ldots, 0)^T. \tag{F.16}$$

Writing matrix $\boldsymbol{U}$ as

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{A}^{(|\mathscr{X}|\times|\mathscr{X}|)} & \boldsymbol{B}^{(|\mathscr{X}|\times 3)} \\ \boldsymbol{C}^{(3\times|\mathscr{X}|)} & \boldsymbol{D}^{(3\times 3)} \end{pmatrix}, \tag{F.17}$$

we have

$$\boldsymbol{A} = (1 + \mu^*)\mathrm{diag}\Big(\frac{1}{Q_{\mu^*}(1)}, \ldots, \frac{1}{Q_{\mu^*}(|\mathscr{X}|)}\Big), \tag{F.18}$$

$$\boldsymbol{B} = \begin{pmatrix} 1 & 1 + \log \frac{Q_{\mu^*}(1)}{P_0(1)} & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 + \log \frac{Q_{\mu^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)} & 0 \end{pmatrix}, \ \boldsymbol{C} = \begin{pmatrix} 1 & \ldots & 1 \\ 1 + \log \frac{Q_{\mu^*}(1)}{P_0(1)} & \ldots & 1 + \log \frac{Q_{\mu^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)} \\ 1 + \log \frac{Q_{\mu^*}(1)}{Q_1(1)} & \ldots & 1 + \log \frac{Q_{\mu^*}(|\mathscr{X}|)}{Q_1(|\mathscr{X}|)} \end{pmatrix}, \ \boldsymbol{D} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{F.19}$$

Also writing $\boldsymbol{S}$ as

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{F}^{(|\mathscr{X}|\times|\mathscr{X}|)} \\ \boldsymbol{K}^{(3\times|\mathscr{X}|)} \end{pmatrix}, \tag{F.20}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

we have

$$\boldsymbol{F} = \operatorname{diag}\left(\frac{1}{Q_1(1)}, \ldots, \frac{1}{Q_1(|\mathscr{X}|)}\right), \tag{F.21}$$

$$\boldsymbol{K} = \begin{pmatrix} 0 & \ldots & 0 \\ 0 & \ldots & 0 \\ \frac{Q_{\mu^*}(1)}{Q_1(1)} & \ldots & \frac{Q_{\mu^*}(|\mathscr{X}|)}{Q_1(|\mathscr{X}|)} \end{pmatrix}. \tag{F.22}$$

By the blockwise inversion formula, we have

$$\boldsymbol{U}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{BMCA}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{BM} \\ -\boldsymbol{MCA}^{-1} & \boldsymbol{M} \end{pmatrix}, \tag{F.23}$$

where $\boldsymbol{M} = \big(\boldsymbol{D} - \boldsymbol{CA}^{-1}\boldsymbol{B}\big)^{-1}$. Since we are only interested in $\frac{\partial Q_{\mu^*}}{\partial Q_1}$, it suffices to find the first $|\mathscr{X}|$ rows of $\boldsymbol{U}^{-1}$. Applying the block inversion formula to find $\boldsymbol{M}$, we get

$$\boldsymbol{BM} = \frac{1+\mu^*}{\operatorname{Var}_{Q_{\mu^*}}\big(\log\frac{Q_{\mu^*}}{P_1}\big)} \begin{pmatrix} (1+E_0)\big(\log\frac{Q_{\mu^*}(1)}{P_0(1)} - E_0\big) - \operatorname{Var}_{Q_{\mu^*}}\big(\log\frac{Q_{\mu^*}}{P_0}\big), & -\big(\log\frac{Q_{\mu^*}(1)}{P_0(1)} - E_0\big), & 0 \\ \vdots & \vdots & \vdots \\ (1+E_0)\big(\log\frac{Q_{\mu^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)} - E_0\big) - \operatorname{Var}_{Q_{\mu^*}}\big(\log\frac{Q_{\mu^*}}{P_0}\big), & -\big(\log\frac{Q_{\mu^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)} - E_0\big), & 0 \end{pmatrix}, \tag{F.24}$$

where we used that at the optimal solution $D(Q_{\mu^*}\|P_0) = E_0$. Furthermore, we get

$$\boldsymbol{A}^{-1}\boldsymbol{BMCA}^{-1} = -\frac{1}{(1+\mu^*)}(\boldsymbol{Q} + \boldsymbol{W}). \tag{F.25}$$

Finally, by the structure of $\boldsymbol{K}$ and $\boldsymbol{BM}$, we see that

$$-\operatorname{diag}\Big(\frac{1}{Q_1}\Big) * \Big(\frac{\partial Q_{\mu^*}}{\partial Q_1}\Big)\Big|_{Q_1=P_1} = -\boldsymbol{F}\big(\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{BMCA}^{-1}\big)\boldsymbol{F} \tag{F.26}$$

$$= \frac{1}{1+\mu^*}\boldsymbol{J}\Big[\boldsymbol{T} - (\boldsymbol{Q} + \boldsymbol{W})\Big]\boldsymbol{J}. \tag{F.27}$$

Substituting (F.26) into (F.4), we get the Hessian matrix in (C.5).

G. Proof of Lemma C.2

Similarly to Lemma C.1 by writing the Lagrangian and using the envelope theorem [27], we have

$$L(Q, Q_1, \eta, \nu) = D(Q\|P_1) + \eta\big(D(Q\|P_0) - \beta D(Q\|Q_1) + \gamma(E_0, Q_1)\big) + \nu\Big(\sum_{x\in\mathscr{X}} Q(x) - 1\Big), \tag{G.1}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

hence

$$\frac{\partial E_1(E_0,Q_1)}{\partial Q_1(j)}=\frac{\partial L}{\partial Q_1(j)}\bigg|_{Q=\hat{Q}_{\eta^*}}=\eta^*\left(\beta\frac{\hat{Q}_{\eta^*}(j)}{Q_1(j)}+\frac{\partial\gamma}{\partial Q_1(j)}\right),\tag{G.2}$$

$$\hat{Q}_{\eta^*}(x)=\frac{P_1^{\frac{1}{1+\eta^*-\eta^*\beta}}(x)Q_1^{\frac{-\eta^*\beta}{1+\eta^*-\eta^*\beta}}(x)P_0^{\frac{\eta^*}{1+\eta^*-\eta^*\beta}}(x)}{\sum_{a\in\mathscr{X}}P_1^{\frac{1}{1+\eta^*-\eta^*\beta}}(a)Q_1^{\frac{-\eta^*\beta}{1+\eta^*-\eta^*\beta}}(a)P_0^{\frac{\eta^*}{1+\eta^*-\eta^*\beta}}(a)}.\tag{G.3}$$

Letting $Q_1=P_1$, we get

$$Q_{\eta^*_\beta}(x)=\frac{P_1^{\frac{1-\eta^*_\beta\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(x)P_0^{\frac{\eta^*_\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(x)}{\sum_{a\in\mathscr{X}}P_1^{\frac{1-\eta^*_\beta\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(a)P_0^{\frac{\eta^*_\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}}(a)}.\tag{G.4}$$

Furthermore, since $D(Q_{\eta^*_\beta}\|P_0)=E_0$, it is easy to see that, $Q_{\eta^*_\beta}=Q_{\lambda^*}$ in (2.11), and $\frac{\eta^*_\beta}{1+\eta^*_\beta-\eta^*_\beta\beta}=\lambda^*$. Also, by (B.8), $Q_{\eta^*_\beta}=Q_{\mu^*}$, and finally from (C.2), we conclude that

$$\left(\beta\frac{\hat{Q}_{\eta^*}(j)}{Q_1(j)}+\frac{\partial\gamma}{\partial Q_1(j)}\right)\bigg|_{Q_1=P_1}=\beta\frac{Q_{\eta^*_\beta}(j)}{P_1(j)}-\beta\frac{Q_{\mu^*}(j)}{P_1(j)}\tag{G.5}$$

$$=0,\tag{G.6}$$

which concludes (C.9). Next, by taking the second derivative of (H.2), we get

$$\frac{1}{\beta}\frac{\partial^2E_1(E_0,Q_1)}{\partial Q_1(i)Q_1(j)}=\frac{\partial\eta^*}{\partial Q_1(i)}\left(\frac{\hat{Q}_{\eta^*}(j)}{Q_1(j)}+\frac{1}{\beta}\frac{\partial\gamma}{\partial Q_1(j)}\right)\tag{G.7}$$

$$+\eta^*\left(\frac{1}{Q_1(j)}\frac{\partial\hat{Q}_{\eta^*}(j)}{\partial Q_1(i)}-\frac{\hat{Q}_{\eta^*}(i)}{Q_1^2(i)}\mathbb{1}\{i=j\}+\frac{1}{\beta}\frac{\partial^2\gamma}{\partial Q_1(i)Q_1(j)}\right).\tag{G.8}$$

Letting $Q_1=P_1$, the first term is zero, hence we only require to find $\frac{\partial\hat{Q}_{\eta^*}(j)}{\partial Q_1(i)}\Big|_{Q_1=P_1}$. For simplicity in our derivations, we write all the expressions in terms of the optimal Lagrange multipliers when $\beta=1$. Since for every $0<\beta\le1$, the optimizing distribution Q when $Q_1=P_1$ is the tilted distribution of P_0, and P_1, and since for every such Q, the optimizing distribution should satisfy the condition (B.8), hence the tilted exponent is equal for every β and by equating the exponents in (H.4) we can define

$$\rho\triangleq\frac{\eta^*_\beta}{\eta^*_1}=1+\eta^*_\beta-\beta\eta^*_\beta,\tag{G.9}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where η_β^* is the Lagrange multiplier for aribitrary β and η_1^* is the Lagrange multiplier when $\beta = 1$. Using proposition F.1 and letting $x = Q$, and $a = Q_1$, and setting $Q_1 = P_1$, we have

$$\boldsymbol{F}_{QQ}^{\beta} = \rho \boldsymbol{F}_{QQ}, \quad \boldsymbol{F}_{QQ_1}^{\beta} = \rho \boldsymbol{F}_{QQ_1}, \tag{G.10}$$

$$\boldsymbol{F}_{Q}^{\beta} = \boldsymbol{F}_{Q}, \quad \boldsymbol{F}_{Q_1}^{\beta} = \boldsymbol{B}_{Q}, \tag{G.11}$$

$$\boldsymbol{G}_{Q}^{\beta} = \beta \boldsymbol{G}_{Q} + (1-\beta)\boldsymbol{F}_{Q}, \quad \boldsymbol{G}_{Q_1}^{\beta} = \boldsymbol{G}_{Q_1}, \tag{G.12}$$

$$\boldsymbol{H}_{Q}^{\beta} = \boldsymbol{H}_{Q} \quad \boldsymbol{H}_{Q_1}^{\beta} = \boldsymbol{H}_{Q}, \tag{G.13}$$

where

$$\boldsymbol{F}_{QQ} = \operatorname{diag}\left(\frac{1}{Q_{\eta_1^*}(1)}, \ldots, \frac{1}{Q_{\eta_1^*}(|\mathscr{X}|)}\right), \tag{G.14}$$

$$\boldsymbol{F}_{QQ_1} = \lambda_1^* \operatorname{diag}\left(\frac{1}{P_1(1)}, \ldots, \frac{1}{P_1(|\mathscr{X}|)}\right), \tag{G.15}$$

$$\boldsymbol{F}_{Q} = \left(1 + \log \frac{Q_{\eta_1^*}(1)}{P_1(1)}, \ldots, 1 + \log \frac{Q_{\eta_1^*}(|\mathscr{X}|)}{P_1(|\mathscr{X}|)}\right)^T, \tag{G.16}$$

$$\boldsymbol{F}_{Q_1} = (0, \ldots, 0)^T, \tag{G.17}$$

$$\boldsymbol{B}_{Q} = \left(1 + \log \frac{Q_{\eta_1^*}(1)}{P_0(1)}, \ldots, 1 + \log \frac{Q_{\eta_1^*}(|\mathscr{X}|)}{P_0(|\mathscr{X}|)}\right)^T, \tag{G.18}$$

$$\boldsymbol{G}_{Q} = \left(\log \frac{P_1(1)}{P_0(1)}, \ldots, \log \frac{P_1(|\mathscr{X}|)}{P_0(|\mathscr{X}|)}\right)^T, \tag{G.19}$$

$$\begin{aligned} \boldsymbol{G}_{Q_1} &= \left(\left(\frac{\hat{Q}_{\eta_1^*}(1)}{Q_1(1)} + \frac{\partial \gamma}{\partial Q_1(1)}\right)\Bigg|_{Q_1=P_1}, \ldots, \left(\frac{\hat{Q}_{\eta_1^*}(|\mathscr{X}|)}{Q_1(|\mathscr{X}|)} + \frac{\partial \gamma}{\partial Q_1(|\mathscr{X}|)}\right)\Bigg|_{Q_1=P_1}\right)^T && \text{(G.20)} \\ &= (0, \ldots, 0)^T, && \text{(G.21)} \end{aligned}$$

$$\boldsymbol{H}_{Q} = (1, \ldots, 1)^T, \tag{G.22}$$

$$\boldsymbol{H}_{Q_1} = (0, \ldots, 0)^T. \tag{G.23}$$

Similarly to the Lemma C.1, we can write $\boldsymbol{U}$ as

$$\boldsymbol{U}_\beta = \begin{pmatrix} \boldsymbol{A}_\beta^{(|\mathscr{X}| \times |\mathscr{X}|)} & \boldsymbol{B}_\beta^{(|\mathscr{X}| \times 3)} \\ \boldsymbol{C}_\beta^{(3 \times |\mathscr{X}|)} & \boldsymbol{D}_\beta^{(3 \times 3)} \end{pmatrix}, \tag{G.24}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where

$$\boldsymbol{A}_\beta = \rho \boldsymbol{A}_1 \tag{G.25}$$

$$\boldsymbol{B}_\beta = \beta \boldsymbol{B}_1 + (1-\beta)\boldsymbol{E}_0 \tag{G.26}$$

$$\boldsymbol{C}_\beta = \beta \boldsymbol{C}_1 + (1-\beta)\boldsymbol{L}_0 \tag{G.27}$$

$$\boldsymbol{D}_\beta = \boldsymbol{D}_1 \tag{G.28}$$

and

$$\boldsymbol{A}_1 = \operatorname{diag}\left(\frac{1}{Q_{\eta_1^*}(1)}, \dots, \frac{1}{Q_{\eta_1^*}(\left|\mathscr{X}\right|)}\right), \quad \boldsymbol{D}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \tag{G.29}$$

$$\boldsymbol{B}_1 = \begin{pmatrix} 1 & \log \frac{P_1(1)}{P_0(1)} & 0 \\ \vdots & \vdots & \vdots \\ 1 & \log \frac{P_1(\left|\mathscr{X}\right|)}{P_0(\left|\mathscr{X}\right|)} & 0 \end{pmatrix}, \quad \boldsymbol{E}_0 = \begin{pmatrix} 1 & 1 + \log \frac{Q_{\eta_1^*}(1)}{P_0(1)} & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 + \log \frac{Q_{\eta_1^*}}{P_0(\left|\mathscr{X}\right|)} & 0 \end{pmatrix}, \tag{G.30}$$

$$\mathit{quad} \tag{G.30}$$

$$\boldsymbol{C}_1 = \begin{pmatrix} 1 & \dots & 1 \\ \log \frac{P_1(1)}{P_0(1)} & \dots & \log \frac{P_1(\left|\mathscr{X}\right|)}{P_0(\left|\mathscr{X}\right|)} \\ 1 + \log \frac{Q_{\eta_1^*}(1)}{P_1(1)} & \dots & 1 + \log \frac{Q_{\eta_1^*}(\left|\mathscr{X}\right|)}{P_1(\left|\mathscr{X}\right|)} \end{pmatrix}, \quad \boldsymbol{L}_0 = \begin{pmatrix} 1 & \dots & 1 \\ 1 + \log \frac{Q_{\eta_1^*}(1)}{P_0(1)} & \dots & 1 + \log \frac{Q_{\eta_1^*}(1)}{P_0(\left|\mathscr{X}\right|)} \\ 1 + \log \frac{Q_{\eta_1^*}(1)}{P_1(1)} & \dots & 1 + \log \frac{Q_{\eta_1^*}(\left|\mathscr{X}\right|)}{P_1(\left|\mathscr{X}\right|)} \end{pmatrix}. \tag{G.31}$$

Also writing $\boldsymbol{S}$ as

$$\boldsymbol{S} = \begin{pmatrix} -\boldsymbol{F}^{\beta}_{QQ_1} \\ \boldsymbol{K} \end{pmatrix}, \tag{G.32}$$

we have

$$\boldsymbol{K}^{(3\times\left|\mathscr{X}\right|)} = \begin{pmatrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & 0 \end{pmatrix}. \tag{G.33}$$

By block inversion formula in (F.23), and using similar arguments as in Lemma C.1, we have

$$\boldsymbol{A}_\beta^{-1}\boldsymbol{B}_\beta\left(\boldsymbol{D}_\beta - \boldsymbol{C}_\beta\boldsymbol{A}_\beta^{-1}\boldsymbol{B}_\beta\right)^{-1}\boldsymbol{C}_\beta\boldsymbol{A}_\beta^{-1} = -\frac{1}{\rho}\left(\boldsymbol{T}^2 + \boldsymbol{V}\right), \tag{G.34}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where $\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^T$ and

$$\boldsymbol{v} = \frac{1}{\sqrt{\mathrm{Var} Q_{\eta^*_\beta}(\Omega)}} \Big(Q_{\eta^*_\beta}(1)\Omega(1), \ldots, Q_{\eta^*_\beta}(|\mathscr{X}|)\Omega(|\mathscr{X}|)\Big), \tag{G.35}$$

$$\Omega(i) = \beta \log \frac{P_1(i)}{P_0(i)} + (1-\beta) \log \frac{Q_{\eta^*_\beta}(i)}{P_0(i)}, i \in \mathscr{X}. \tag{G.36}$$

Finally, by the structure of $\boldsymbol{S}$, we find

$$\mathrm{diag}\Big(\frac{1}{P_1}\Big) * \Big(\frac{\partial \hat{Q}_{\eta^*}}{\partial Q_1}\Big)\Big|_{Q_1=P_1} = \boldsymbol{J}\Big[-\boldsymbol{T} + (\boldsymbol{T}^2 + \boldsymbol{V})\Big]\boldsymbol{J}. \tag{G.37}$$

Substituting (H.37) and (C.5) into (H.7), we get

$$\boldsymbol{H}_2 = \rho\eta_1^*\boldsymbol{J}\Big[-\boldsymbol{T} + \boldsymbol{Q} + \eta_1^*\boldsymbol{V} + (1-\eta_1^*)\boldsymbol{W}\Big]\boldsymbol{J}, \tag{G.38}$$

where we have used the identity $\eta_1^* = \frac{\mu^*}{1+\mu^*}$ derived by setting the tilted distribution exponents to be equal for when $Q_1 = P_1$.

H. Proof of Lemma C.3

By applying a Taylor expansion to $E_1(E_0, Q_1)$ around $Q_1 = P_1$, we obtain

$$E_1(E_0, Q_1) = E_1(E_0, P_1) + \boldsymbol{\theta}_{P_1}^T \nabla E_1(E_0, Q_1)\big|_{Q_1=P_1} + \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{H}_2 \boldsymbol{\theta}_{P_1} + o(\|\boldsymbol{\theta}_{P_2}\|_\infty^2). \tag{H.1}$$

The first term in the expansion is $E_1^*(E_0)$. Also by Lemma C.2, the gradient evaluated at $Q_1 = P_1$ is zero. Further approximating the constraint in (B.30), we get

$$D(Q_1\|P_1) = \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J} \boldsymbol{\theta}_{P_1} + o(\|\boldsymbol{\theta}_{P_1}\|_\infty^2). \tag{H.2}$$

By substituting the expansions (H1) and (H2) in (B.39), we obtain

$$E_1(E_0, r) = E_1^*(E_0) + \min_{\substack{\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J}\boldsymbol{\theta}_{P_1} + o(\|\boldsymbol{\theta}_{P_1}\|_\infty^2) \le r \\ \mathbf{1}^T\boldsymbol{\theta}_{P_1} = 0}} \frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{H}_2 \boldsymbol{\theta}_{P_1} + o(\|\boldsymbol{\theta}_{P_1}\|_\infty^2). \tag{H.3}$$

To find the error term of the above approximation as a function of r, first observe that we can take $o(\|\boldsymbol{\theta}_{P_1}\|_\infty^2)$ out of minimization and substitute it with $o(\|\boldsymbol{\theta}_{P_1}^*(r)\|_\infty^2)$, where $\boldsymbol{\theta}_{P_1}^*$ is the optimal solution of the minimization. Moreover, approximating the inequality constraint can result an error of $o(\sqrt{r})$ in $\|\boldsymbol{\theta}_{P_1}^*\|_\infty$. Therefore, from the inequality constraint and using the fact that it imposes a constraint on the length of the vector $\boldsymbol{\theta}$, we have that $\|\boldsymbol{\theta}_{P_1}^*\|_\infty \le c\sqrt{r} + o(\sqrt{r})$, where c is independent from r and only depends on $\boldsymbol{J}, \boldsymbol{H}_2$. This argument together with (H3) concludes the proof of (C.14).

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

I. Proof of Lemma C.4

Since $\boldsymbol{J}$ is a diagonal matrix with non-zero diagonal entries, we have

$$\sqrt{\boldsymbol{J}} = \operatorname{diag}\left(\frac{1}{\sqrt{P_1(1)}}, \ldots, \frac{1}{\sqrt{P_1(|\mathscr{X}|)}}\right). \tag{I.1}$$

Letting $\boldsymbol{\psi}_{P_1} = \sqrt{\boldsymbol{J}}\boldsymbol{\theta}_{P_1}$, we obtain

$$\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{H}_2 \boldsymbol{\theta}_{P_1} = \frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1}, \tag{I.2}$$

$$\frac{1}{2}\boldsymbol{\theta}_{P_1}^T \boldsymbol{J} \boldsymbol{\theta}_{P_1} = \frac{1}{2}\boldsymbol{\psi}_{P_1}^T \boldsymbol{\psi}_{P_1}, \tag{I.3}$$

$$\mathbf{1}^T \boldsymbol{\theta}_{P_1} = \mathbf{1}^T \sqrt{\boldsymbol{J}}^{-1} \boldsymbol{\psi}_{P_1}. \tag{I.4}$$

Next we show that we can drop the equality constraint. Note that $\boldsymbol{\sigma}_1 \triangleq \frac{\sqrt{\boldsymbol{J}}^{-1}\mathbf{1}}{\|\sqrt{\boldsymbol{J}}^{-1}\mathbf{1}\|}$ is in the null space of $\boldsymbol{H}$, i.e.

$$\boldsymbol{H}\sqrt{\boldsymbol{J}}^{-1}\mathbf{1} = \beta\eta_\beta^*\sqrt{\boldsymbol{J}}\Big[\boldsymbol{Q} + \eta_1^*\boldsymbol{V} + (1-\eta_1^*)\boldsymbol{W} - \boldsymbol{T}\Big]\mathbf{1} = \mathbf{0} \tag{I.5}$$

since $\boldsymbol{v}^T\mathbf{1} = 0, \boldsymbol{w}^T\mathbf{1} = 0, \boldsymbol{T}\mathbf{1} = \boldsymbol{q}$. Now, assume the optimizer of the second optimization in (C.16) is of the form

$$\boldsymbol{\psi}_{P_1} = \rho_1\boldsymbol{\sigma}_1 + \rho_2\boldsymbol{\sigma}_2, \tag{I.6}$$

where $\boldsymbol{\sigma}_1 \perp \boldsymbol{\sigma}_2$, $\|\boldsymbol{\sigma}_2\| = 1$. Then,

$$\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1} = (\rho_1\boldsymbol{\sigma}_1 + \rho_2\boldsymbol{\sigma}_2)^T \boldsymbol{H} (\rho_1\boldsymbol{\sigma}_1 + \rho_2\boldsymbol{\sigma}_2) \tag{I.7}$$

$$= \rho_2^2 \boldsymbol{\sigma}_2^T \boldsymbol{H} \boldsymbol{\sigma}_2, \tag{I.8}$$

where we also used $\boldsymbol{H}\boldsymbol{\sigma}_1 = 0, \boldsymbol{H} = \boldsymbol{H}^T$. Assuming $\boldsymbol{H}$ has negative eigenvalues (otherwise, both optimization problems are equal to zero), the inequality constraint should be satisfied with equality. Then,

$$\boldsymbol{\psi}_{P_1}^T \boldsymbol{H} \boldsymbol{\psi}_{P_1} = \left(2r - \rho_1^2\right) \boldsymbol{\sigma}_2^T \boldsymbol{H} \boldsymbol{\sigma}_2. \tag{I.9}$$

Therefore, to achieve the minimum, ρ_1 must be zero, which concludes the proof.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

J. Proof of Theorem 4.1

From Theorem 3.1, it can be shown that by taking $E_0 = \Theta(n^{-1})$, there exists an α such that the classifier proposed in (3.1) achieves $E_1 = D(P_0\|P_1)$ which is equal to the Stein regime exponent of hypothesis testing with known distributions. This implies that (4.3) is equal to $D(P_0\|P_1)$. In fact, there is no need for a training sequence to achieve (4.3). Since Hoeffding's test achieves the optimal error exponent tradeoff only by knowing distribution P_0, it is easy to see that the Hoeffding test achieves the Stein regime exponent for the unknown distribution for any P_1. However, in order to achieve the largest error exponent under P_0 that guarantees that for any P_1 the type-II probability of error is bounded by some $\varepsilon \in (0,1)$, Hoeffding's and our proposed classifier are not universal, since for any choice of threshold $E_0 > 0$, the type-II probability of error for any distribution P_1 such that $D(P_1\|P_0) < E_0$ tends to one. We show that Gutman's universal test [16] achieves the largest type-I error exponent, while the type-II probability of error is bounded away from one. Using the Gutman's test, we obtain

$$D_\alpha^{\mathrm{GJS}}(\hat{T}_{\boldsymbol{x}}\|\hat{T}_{\boldsymbol{z}}) \leq \frac{1}{2n} G^{-1}_{|\mathscr{X}|-1}(\varepsilon), \tag{J.1}$$

where

$$D_\alpha^{\mathrm{GJS}}(Q\|P) = D\left(Q\middle\|\frac{Q+\alpha P}{1+\alpha}\right) + \alpha D\left(P\middle\|\frac{Q+\alpha P}{1+\alpha}\right), \tag{J.2}$$

is the generalized Jensen–Shannon divergence and $G_a^{-1}(.)$ is the inverse of the complementary CDF of a chi-squared random variable with a degrees of freedom; [42] shows that Gutman's test with the chosen threshold achieves type-II error probability of ε for any P_1 and achieves the type-I error exponent $D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$. We prove the converse showing that the achievable error exponent using Gutman's test is, in fact, the Stein's regime exponent defined in (4.3). Note that as $\alpha \to \infty$, i.e. when the number of training samples is much larger than the number of test samples, then the exponent $D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$ converges to $D(P_1\|P_0)$ which is the Stein regime error exponent under when both distributions are known. We prove the asymptotic optimality of Gutman's test in this setting. Specifically, we show that for any test such that the type-I error exponent $E_0^{(\varepsilon)} > D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$, there exists a P_1 such that type-II probability of error of the test tends to one, i.e. for any test ϕ_n such that

$$\limsup_{n\to\infty} \varepsilon_1(\phi_n) \leq \varepsilon \tag{J.3}$$

for all $P_1 \in \mathscr{P}(\mathscr{X})$ then

$$\lim_{n\to\infty} E_0^{(\varepsilon)}(\phi_n) \leq D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0). \tag{J.4}$$

First, by [42, Lemma 6], any optimal test can be converted into a test based on the type of the observation samples, and the error probabilities of such a test only change by a constant factor. Hence, the converted type-based test is asymptotically optimal, and we can limit the classifiers to type-based ones when studying the error exponent. Next, using the similar idea to [42], we show that in order to have a type-II error probability bounded away form one for every P_1, the classifier necessarily needs to decide in favor of the second hypothesis when $\|\hat{T}_{\boldsymbol{x}} - \hat{T}_{\boldsymbol{z}}\| \leq \delta$, for $\delta > 0$, since under P_1 the types of the test and training sequence will converge to P_1. In addition, we only consider deterministic tests, since

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

it can be shown that randomizing the test cannot increase the error exponents. We have the following lemma.

LEMMA J.1 Let $\varepsilon \in (0,1)$ and $\boldsymbol{X}, \boldsymbol{x}$ be two independent i.i.d sequences, generated by distribution P. Then,

$$P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-\hat{T}_{\boldsymbol{x}}(a)\big|\geq\varepsilon\Big)\leq 2\big|\mathscr{X}\big|e^{-n\frac{\varepsilon^2}{2}}+2\big|\mathscr{X}\big|e^{-\alpha n\frac{\varepsilon^2}{2}}. \tag{J.5}$$

Proof. By the triangle inequality, union bound and Hoeffding's inequality [4], we have

$$\begin{aligned}
P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-\hat{T}_{\boldsymbol{x}}(a)\big|\geq\varepsilon\Big) &= P\big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-P(a)+P(a)-\hat{T}_{\boldsymbol{x}}(a)\big|\geq\varepsilon\Big) && \text{(J.6)}\\
&\leq P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-P(a)\big|+\big|\hat{T}_{\boldsymbol{x}}(a)-P(a)\big|\geq\varepsilon\Big) && \text{(J.7)}\\
&\leq P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-P(a)\big|+\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{x}}(a)-P(a)\big|\geq\varepsilon\Big) && \text{(J.8)}\\
&\leq P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-P(a)\big|\geq\frac{\varepsilon}{2}\cup\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{x}}(a)-P(a)\big|\geq\frac{\varepsilon}{2}\Big) && \text{(J.9)}\\
&\leq P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{z}}(a)-P(a)\big|\geq\frac{\varepsilon}{2}\Big)+P\Big(\max_{a\in\mathscr{X}}\big|\hat{T}_{\boldsymbol{x}}(a)-P(a)\big|\geq\frac{\varepsilon}{2}\Big) && \text{(J.10)}\\
&\leq\sum_{a\in\mathscr{X}}P\Big(\Big|\hat{T}_{\boldsymbol{z}}(a)-P(a)\Big|\geq\frac{\varepsilon}{2}\Big)+\sum_{a\in\mathscr{X}}P\Big(\Big|\hat{T}_{\boldsymbol{x}}(a)-P(a)\Big|\geq\frac{\varepsilon}{2}\Big) && \text{(J.11)}\\
&\leq 2\big|\mathscr{X}\big|e^{-2n(\frac{\varepsilon}{2})^2}+2\big|\mathscr{X}\big|e^{-2k(\frac{\varepsilon}{2})^2}, && \text{(J.12)}
\end{aligned}$$

which concludes the proof of the lemma. □

LEMMA J.2 Let $(Q,Q_1)\in\mathscr{P}_n(\mathscr{X})\times\mathscr{P}_k(\mathscr{X})$ satisfy

$$\max_{a\in\mathscr{X}}\Big\{\Big|Q(a)-Q_1(a)\Big|\Big\}\leq\sqrt{\frac{2}{(\alpha\wedge 1)n}\log\frac{4\big|\mathscr{X}\big|}{1-\varepsilon}}, \tag{J.13}$$

where $a\wedge b=\min\{a,b\}$. Then for any type-based test $\phi_n(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})$ such that for all distributions $\tilde{P}_1\in\mathscr{P}(\mathscr{X})$

$$\varepsilon_1\big(\phi_n(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})\big)<\varepsilon,\quad \varepsilon\in(0,1), \tag{J.14}$$

we have $\phi_n(Q,Q_1)=1$.

Proof. We prove this lemma by contradiction. Assume there exists a type-based test ϕ_n such that $\varepsilon_1\big(\phi_n(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})\big)<\varepsilon$, and $\phi_n(Q,Q_1)=0$ for types Q, Q_1 satisfuing (J.13). Then the type-II probability of error for such test will be such that

$$\varepsilon_1(\phi_n)=P_1(\phi_n(\hat{T}_{\boldsymbol{x}},\hat{T}_{\boldsymbol{z}})=0) \tag{J.15}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

$$\geq P_1(\phi_n(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) = 0, \hat{T}_{\boldsymbol{x}} = Q, \hat{T}_{\boldsymbol{z}} = Q_1). \tag{J.16}$$

Conditioning on training and test sequences, we obtain

$$\varepsilon_1\big(\phi_n\big) \geq P_1(\phi_n(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) = 0 \big| \hat{T}_{\boldsymbol{x}} = Q, \hat{T}_{\boldsymbol{z}} = Q_1\big) P_1(\hat{T}_{\boldsymbol{x}} = Q, \hat{T}_{\boldsymbol{z}} = Q_1\big) \tag{J.17}$$

$$= P_1(\hat{T}_{\boldsymbol{x}} = Q, \hat{T}_{\boldsymbol{z}} = Q_1\big) \tag{J.18}$$

$$\geq \left(1 - \frac{(1-\varepsilon)}{2} - \frac{(1-\varepsilon)}{2}\right) \tag{J.19}$$

$$\geq \varepsilon, \tag{J.20}$$

where in the last step we used the Lemma J.1. Therefore, for any probability distribution P_1 the type-II error probability exceeds ε, contradicting the initial assumption. Hence, any type-based classifier such that $\varepsilon_1(\phi_n) < \varepsilon$ for all distributions of $\tilde{P}_1 \in \mathscr{P}(\mathscr{X})$ should satisfy $\phi(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) = 1$ for sufficiently close types $\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}$. □

Finally by the method of types, we can lower bound the type-I probability of error as

$$\varepsilon_0 = P_0(\phi_n(Q, Q_1) = 1) \tag{J.21}$$

$$\geq (n+1)^{-|\mathscr{X}|}(k+1)^{-|\mathscr{X}|} \sum_{\phi_n(Q,Q_1)=1} e^{-n\left(D(Q\|P_0)+\alpha D(Q_1\|P_1)\right)} \tag{J.22}$$

$$\geq (n+1)^{-|\mathscr{X}|}(k+1)^{-|\mathscr{X}|} e^{-n\left(\min_{\phi_n(Q,Q_1)=1} D(Q\|P_0)+\alpha D(Q_1\|P_1)\right)}. \tag{J.23}$$

Therefore, for any type-based test, the type-I error exponent is upper bounded by

$$E_0(\phi) \leq \liminf_{n\to\infty} \min_{\phi_n(Q,Q_1)=1} D(Q\|P_0) + \alpha D(Q_1\|P_1). \tag{J.24}$$

Now by Lemma J.2, for any test with type-II error probability bounded away from one, we have

$$E_0(\phi) \leq \liminf_{n\to\infty} \min_{\max_{a\in\mathscr{X}}|Q(a)-Q_1(a)| \leq \sqrt{\frac{2}{(\alpha\wedge 1)n}\log\frac{4|\mathscr{X}|}{1-\varepsilon}}} D(Q\|P_0) + \alpha D(Q_1\|P_1) \tag{J.25}$$

$$= \lim_{n\to\infty} \min_{Q\in\mathscr{P}(\mathscr{X})} D(Q\|P_0) + \alpha D(Q_1\|P_1) + o(1) \tag{J.26}$$

$$= \min_{Q\in\mathscr{P}(\mathscr{X})} D(Q\|P_0) + \alpha D(Q\|P_1) \tag{J.27}$$

$$= D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0), \tag{J.28}$$

where the last step follows from, e.g. [36]. This concludes the proof. Observe that the proof can be easily generalized for the case where both probability distributions are unknown and only training samples

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

from both are given, i.e. the largest type-I error exponent achievable when the type-II error probability is bounded away from one is also $D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$ and can be attained by Gutman's test.

K. Proof of Theorem 5.1

We first find the error probabilities as a function of thresholds, and then, we find the average stopping time under each hypothesis. The type-I error probability can be upper bounded by

$$\varepsilon_0 \le \sum_{t=n}^{\infty} \mathbb{P}\Big[t\big(D(\hat{T}_{\boldsymbol{x}}\|P_0) - D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_{\boldsymbol{z}})\big) \ge \gamma_{1,n}(t)\Big]. \tag{K.1}$$

By the method of types [11], we have

$$\varepsilon_0 \le \sum_{t=n}^{\infty} \sum_{(Q,Q_1)\in\mathscr{Q}_{01}(t)\cap\mathscr{P}_t(\mathscr{X})\times\mathscr{P}_{\alpha t}(\mathscr{X})} e^{-t\big(D(Q\|P_0)+\alpha D(Q_1\|P_1)\big)} \tag{K.2}$$

$$\le \sum_{t=n}^{\infty}(\alpha t+1)^{\left|\mathscr{X}\right|}(t+1)^{\left|\mathscr{X}\right|}e^{-E_{01}(t)}, \tag{K.3}$$

where

$$E_{01}(t) = \min_{(Q,Q_1)\in\mathscr{Q}_{01}(t)} t\big(D(Q\|P_0) + \alpha D(Q_1\|P_1)\big), \tag{K.4}$$

$$\mathscr{Q}_{01}(t) = \left\{(Q,Q_1) : D(Q\|P_0) - D(Q\|Q_1') \ge \frac{\gamma_{1,n}(t)}{t}, Q_1' = (1-\delta_n)Q_1 + \delta_n U\right\}. \tag{K.5}$$

Similarly to the proof of Theorem 3.1, we can expand all the exponents defined in this proof around Q_1, and by choosing $\delta_n = o(n^{-1})$, the error term of the expansion vanishes as n tends to infinity and we can substitutde $Q_1{\prime}$ with Q_1 for $Q_1 \in \mathscr{P}_{\delta_n}(\mathscr{X})$. This is also true for all the exponent functions we define in the rest of the proof. For every fixed Q_1 we can use the dual form of the optimization (K.4) over Q to get [3]

$$E_{01}(t) = \min_{Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}\bigg(\Big(\max_{\lambda\ge 0}\gamma_{1,n}(t)\lambda - t\log\sum_{x\in\mathscr{X}}P_0^{1-\lambda}(x)Q_1^{\lambda}(x)\Big) + \alpha t D(Q_1\|P_1)\bigg). \tag{K.6}$$

Substituting $\gamma_{1,n}(t) = nD(Q_1\|P_0) + (4\left|\mathscr{X}\right|+4)\log(t+1)$ and setting $\lambda = 1$, we obtain the lower bound

$$E_{01}(t) \ge (4\left|\mathscr{X}\right| + 4)\log(t+1) + n\min_{Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})} D(Q_1\|P_0) + \frac{\alpha t}{n}D(Q_1\|P_1), \tag{K.7}$$

where for large enough n

$$E_{01}(t) \ge (4\left|\mathscr{X}\right| + 4)\log(t+1) + nD_{\beta(t)}(P_1\|P_0), \tag{K.8}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

and

$$\beta(t) = \frac{\frac{\alpha t}{n}}{1+\frac{\alpha t}{n}}. \tag{K.9}$$

Furthermore, as $\beta(t)$ is strictly increasing function in t and $D_{\beta(t)}(P_1\|P_0)$ is a non-decreasing function in β [36], then for all $t \geq n$, we have

$$D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0) \leq D_{\beta(t)}(P_1\|P_0). \tag{K.10}$$

Therefore, by (K.8), (K.10)

$$\varepsilon_0 \leq \sum_{t=n}^{\infty}(\alpha t+1)^{\left|\mathscr{X}\right|}(t+1)^{-3\left|\mathscr{X}\right|-4}e^{-nD_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)} \tag{K.11}$$

$$\leq ce^{-nD_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)}, \tag{K.12}$$

where c is a positive constant.

Next, we find a lower bound to the type-II error exponent. Upper bounding the type-II error probability, we have

$$\varepsilon_1 \leq \sum_{t=n}^{\infty}\mathbb{P}\Big[t(D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_{\boldsymbol{z}}) - D(\hat{T}_{\boldsymbol{x}}\|P_0)) \geq \gamma_{0,n}(t)\Big]. \tag{K.13}$$

By the method of types, we have

$$\varepsilon_1 \leq \sum_{t=n}^{\infty}\sum_{(Q,Q_1)\in\mathscr{Q}_{10}(t)\cap\mathscr{P}_t(\mathscr{X})\times\mathscr{P}_{\alpha t}(\mathscr{X})} e^{-t\left(D(Q\|P_1)+\alpha D(Q_1\|P_1)\right)} \tag{K.14}$$

$$\leq \sum_{t=n}^{\infty}(\alpha t+1)^{\left|\mathscr{X}\right|}(t+1)^{\left|\mathscr{X}\right|}e^{-E_{10}(t)}, \tag{K.15}$$

where

$$E_{10}(t) = \min_{(Q,Q_1)\in\mathscr{Q}_{10}(t)} t\big(D(Q\|P_1)+\alpha D(Q_1\|P_1)\big), \tag{K.16}$$

$$\mathscr{Q}_{10}(\gamma) = \left\{(Q,Q_1) : D(Q\|Q_1') - D(Q\|P_0) \geq \frac{\gamma_{0,n}(t)}{t}, Q_1' = (1-\delta_n)Q_1 + \delta_n U\right\}. \tag{K.17}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

We show that there exists a finite α^*_{seq} such that for every $\alpha \geq \alpha^*_{\text{seq}}$, the achievable type-II error exponent is lower bounded by $nD(P_0\|P_1)$. Similarly to the fixed sample sized case, for every $Q_1 \in \mathscr{P}_{\delta_n}(\mathscr{X})$, let

$$E_{10}(Q_1, t) = \min_{\substack{D(Q\|Q_1)-D(Q\|P_0)\geq \frac{\gamma_{0,n}(t)}{t} \\ Q\in\mathscr{P}(\mathscr{X})}} tD(Q\|P_1), \tag{K.18}$$

which is the error exponent when the type of the training sequence is Q_1. We also define

$$E_{10}(r, t) = \min_{\substack{D(Q_1\|P_1)\leq r \\ Q_1\in\mathscr{P}_{\delta_n}(\mathscr{X})}} E_{10}(Q_1, t). \tag{K.19}$$

Next, we expand $E_{10}(r, t)$ using a Taylor expansion. It is sufficient to show that there exists a finite α such that

$$\inf_{\substack{n\leq t \\ t\in\mathbb{N}}} \inf_{0\leq r\leq \frac{r_c}{2}} E_{10}(r, t) + \alpha tr \geq nD(P_1\|P_0) + (2\left|\mathscr{X}\right| + 2)\log(t + 1). \tag{K.20}$$

Equivalently, (K.20) can be written as the following condition:

$$E_{10}(r, t) + \alpha tr \geq nD(P_1\|P_0) + (2\left|\mathscr{X}\right| + 2)\log(t + 1)\ \forall r:\ 0 \leq r \leq \frac{r_c}{2}, n \leq t, t \in \mathbb{N}. \tag{K.21}$$

Using a Taylor series expansion of $E_{10}(r, t)$ around $r = 0$, we have [3]

$$E_{10}(r, t) \geq E_{10}(r = 0, t) + z_1(t)\sqrt{r}\mathbb{1}\{t \geq n + 1\} + hr\mathbb{1}\{t = n\}, \tag{K.22}$$

where

$$z_1(t) = \inf_{0\leq r\leq \frac{r_c}{2}} \frac{\partial E_{10}(r, t)}{\partial\sqrt{r}} \tag{K.23}$$

$$= \inf_{D(Q_1\|P_1)\leq \frac{r_c}{2}} -\sqrt{\operatorname{Var}_{P_1}\left(\lambda^*(t)\frac{Q_{\lambda^*(t)} - \frac{n}{t}P_0}{Q_1}\right)}, \tag{K.24}$$

$$h = \frac{1}{2}\inf_{0\leq r\leq \frac{r_c}{2}} \frac{\partial^2 E_{10}(r, t = n)}{(\partial\sqrt{r})^2}, \tag{K.25}$$

where $Q_{\lambda^*(t)}, \lambda^*(t)$ are the minimizing distribution and the Lagrange multiplier in (K.18). We have used the fact that for $t = n$, the optimization problem $E_{10}(r, t = n)$ is the same as the fixed sample sized classifier and hence we can use the result of the Theorem 3.5 to lower bound the exponent by $hr\mathbb{1}\{t = n\}$ for some finite h. Also for every $n < t$, the Taylor expansion of $E_{10}(r, t)$ has a non-zero first-order term, hence the expansion includes $\sqrt{r}$. Writing $E_{10}(r = 0, t)$ in the dual form, we have the lower bound

$$E_{10}(r = 0, t) \geq \max_{\frac{1}{2}\leq\lambda} \lambda\gamma - \log\sum_{x\in\mathscr{X}} P_0^{\lambda}(x)P_1^{1-\lambda}(x), \tag{K.26}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where $\gamma = \frac{\gamma_{0,n}(t)}{t}\Big|_{Q_1=P_1}$. Then, $E_{10}(r=0,t)$ is convex as it is the supremum of linear functions in γ [5]. Hence, we can lower bound $E_{10}(r=0,t)$ by expanding it around $\gamma = \frac{nD(P_0\|P_1)}{t}$. Using the envelope theorem, we get [27]

$$E_{10}(r=0,t) \geq \tilde{E}_{10}(r=0,t) + (2\big|\mathscr{X}\big| + 2)\log(t+1), \tag{K.27}$$

where

$$\tilde{E}_{10}(r=0,t) = t\max_{\frac{1}{2}\leq\lambda} \lambda\frac{nD(P_0\|P_1)}{t} - \log\sum_{x\in\mathscr{X}} P_0^{\lambda}(x)P_1^{1-\lambda}(x). \tag{K.28}$$

Further expanding $\tilde{E}_{10}(r=0,t)$ around $t=n$, we have

$$\begin{aligned}E_{10}(r,t) &\geq nD(P_1\|P_0) + (2\big|\mathscr{X}\big| + 2)\log(t+1)\\ &\quad + m(t-n)\mathbb{1}\{t>2n\} + m(t-n)^2\mathbb{1}\{t\leq 2n\} + z_1(t)\sqrt{r}\mathbb{1}\{t\geq n+1\} + hr\mathbb{1}\{t=n\},\end{aligned} \tag{K.29}$$

where we have used the fact that $\frac{\partial\tilde{E}_{10}(r=0,t)}{\partial t}\Big|_{t=n} = 0$, and

$$m = \min\left\{\inf_{2n<t}\frac{\partial\tilde{E}_{10}(r=0,t)}{\partial t}, \frac{1}{2}\inf_{n\leq t\leq 2n}\frac{\partial^2\tilde{E}_{10}(r=0,t)}{\partial t^2}\right\}. \tag{K.30}$$

We have expanded $\tilde{E}_{10}(r=0,t)$ as above since it behaves linearly as t tends to infinity while quadratically for t close to n. Using Proposition F.1, we have

$$\frac{\partial\tilde{E}_{10}(r=0,t)}{\partial t} = -\log\sum_{x\in\mathscr{X}} P_0^{\tilde{\lambda}^*(t)}(x)P_1^{1-\tilde{\lambda}^*(t)}(x), \tag{K.31}$$

$$\frac{\partial^2\tilde{E}_{10}(r=0,t)}{\partial t^2} = \frac{\left(\sum Q^*(t)\log\frac{P_0}{P_1}\right)^2}{t\mathrm{Var}_{P_1}\left(\log\frac{P_0}{P_1}\right)}, \tag{K.32}$$

which are finite and strictly positive for any $2n<t$, $n\leq t\leq 2n$, respectivly, since $Q^*(t)$ is a probability distribution and $\frac{1}{2}\leq\tilde{\lambda}^*(t)<1$ is the optimizer in (K.28). From condition (K.21) by substituting the approximation (K.29), we need that

$$m\Big((t-n)\mathbb{1}\{t>2n\} + (t-n)^2\mathbb{1}\{t\leq 2n\}\Big) + z_1(t)\sqrt{r}\mathbb{1}\{t\geq n+1\} + hr\mathbb{1}\{t=n\} + \alpha tr \geq 0 \tag{K.33}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

for every $0 \le r,\ n \le t, t \in \mathbb{N}$. For $\alpha > \frac{\left|h\right|}{n}$, letting $r^* = \frac{z_1(t)\mathbb{1}\{t \ge n+1\}}{2(\alpha t + h\mathbb{1}\{t=n\})}$ to minimize the LHS of (K.33) over r, we get the condition

$$m\Big((t-n)\mathbb{1}\{t > 2n\} + (t-n)^2\mathbb{1}\{t \le 2n\}\Big) - \frac{\rho(t)}{4(\alpha t + h\mathbb{1}\{t=n\})} \ge 0, \tag{K.34}$$

where

$$\rho(t) = z_1^2(t)\mathbb{1}\{t \ge n+1\} \tag{K.35}$$

and we have used that $0 \le z_1(t) < \infty$ by (K.24) since for every Q_1 satisfying $D(Q_1\|P_1) \le \frac{r_c}{2}$, $\lambda^*(t)$ is finite and $Q_{\lambda^*(t)}$ is a probability distribution, hence the variance in (K.24) is finite. Moreover, $\rho(t)$ equals to zero at $t = n$ and it is finite for every $t \in \mathbb{N}$ with finite limit as $t \to \infty$. Hence, there exists a finite c_1 such that $\rho(t) \le c_1\Big((t-n)\mathbb{1}\{t > 2n\} + (t-n)^2\mathbb{1}\{t \le 2n\}\Big)$ for $t \in \mathbb{N}$. Then, by further relaxing condition (K.33), we need

$$\Big((t-n)\mathbb{1}\{t > 2n\} + (t-n)^2\mathbb{1}\{t \le 2n\}\Big)\Big(m - \frac{c_1}{4\alpha t}\Big) \ge 0, \tag{K.36}$$

where we have dropped the $h\mathbb{1}\{t=n\}$ term as it is only non-zero for $t = n$, which sets $(t-n)$ and $(t-n)^2$ to zero. Therefore, if $\alpha > \frac{c_1}{4m}$, the sufficient condition is satisfied and

$$E_{10}(t) \ge nD(P_0\|P_1) + (2\left|\mathscr{X}\right| + 2)\log(t+1) \tag{K.37}$$

for all $n \le t, t \in \mathbb{N}, 0 \le r \le \frac{r_c}{2}$. Therefore, for $\alpha > \max\Big\{\frac{c_1}{4m}, \left|h\right|, \frac{2D(P_1\|P_0)}{r_c} + 4\left|\mathscr{X}\right| + 4\Big\}$, by substituting (K.37) in (K.15), we get

$$\varepsilon_1 \le ce^{-nD(P_0\|P_1)}, \tag{K.38}$$

where c is a positive constant.

Next, we find the average stopping times of the proposed sequential classifier. We first show the convergence of τ in probability under each hypothesis, and by proving its uniform integrability, we can conclude its convergence in the L^1 norm. The following lemma states that for every n, the classifier stops with probability one.

Lemma K.1 . Let τ_0, τ_1 be the the smallest time that the sequential classifier crosses threshold $\gamma_{0,n}(t)$ or $\gamma_{1,n}(t)$, respectively, i.e.

$$\tau_0 = \inf\{t \ge n : S_t \ge \gamma_{0,n}(t)\}, \qquad \tau_1 = \inf\{t \ge n : S_t \le -\gamma_{1,n}(t)\}. \tag{K.39}$$

Then for $i \in \{0, 1\}$, $t \ge n$,

$$\mathbb{P}_i[\tau_i > t] \le c_i(t+1)^{d_i\left|\mathscr{X}\right|}e^{\xi_i n}e^{-tE_i}, \tag{K.40}$$

where $E_i, \xi_i, c_i, d_i > 0$ and finite.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Proof. By the method of types, the probability of passing the threshold $\gamma_{0,n}(t)$ under the first hypothesis at a time after $t \geq n$ can be upper bounded by

$$\mathbb{P}_0[\tau_0 > t] \leq \mathbb{P}_0\big[S_t \leq \gamma_{0,n}(t)\big] \tag{K.41}$$

$$= \mathbb{P}_0\left[D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_z) - D(\hat{T}_{\boldsymbol{x}}\|P_0) \leq \frac{\gamma_{0,n}(t)}{t}\right] \tag{K.42}$$

$$\leq \sum_{(Q,Q_1)\in\mathcal{Q}_{00}\cap\mathcal{P}_t(\mathcal{X})\times\mathcal{P}_{\alpha t}(\mathcal{X})} e^{-t\left(D(Q\|P_0)+\alpha D(Q_1\|P_1)\right)} \tag{K.43}$$

$$\leq (\alpha t+1)^{\left|\mathcal{X}\right|}(t+1)^{\left|\mathcal{X}\right|}e^{-E_0(t)}, \tag{K.44}$$

where

$$E_0(t) = \min_{(Q,Q_1)\in\mathcal{Q}_{00}(t)} t\big(D(Q\|P_0) + \alpha D(Q_1\|P_1)\big), \tag{K.45}$$

$$\mathcal{Q}_{00}(t) = \left\{(Q,Q_1) : D(Q\|P_0) - D(Q\|Q_1) \geq -\frac{\gamma_{0,n}(t)}{t}\right\}. \tag{K.46}$$

For every fix Q_1, we can use the dual form of the optimization over Q to get [3]

$$E_0(t) = \min_{Q_1\in\mathcal{P}(\mathcal{X})} \left(t \max_{0\leq\lambda} -\frac{\gamma_{0,n}(t)}{t}\lambda - \log\sum_{x\in\mathcal{X}} P_0^{1-\lambda}(x)Q_1^{\lambda}(x)\right) + \alpha t D(Q_1\|P_1)\Big) \tag{K.47}$$

$$\geq \min_{Q_1\in\mathcal{P}(\mathcal{X})} \left(t \max_{0\leq\lambda\leq1} -\frac{\gamma_{0,n}(t)}{t}\lambda - \log\sum_{x\in\mathcal{X}} P_0^{1-\lambda}(x)Q_1^{\lambda}(x)\right) + \alpha t D(Q_1\|P_1)\Big). \tag{K.48}$$

Let

$$E(\gamma) = \max_{0\leq\lambda\leq1} \lambda\gamma - \log\sum_{x\in\mathcal{X}} P_0^{1-\lambda}(x)Q_1^{\lambda}(x). \tag{K.49}$$

Then, $E(\gamma)$ is convex as it is the supremum of linear functions in γ [5]. Therefore, letting the E_0 to be the non-zero minimum of the optimization, we have

$$E\left(\frac{\gamma}{t}\right) \geq E_0 + \frac{\partial E(\gamma)}{\partial\gamma}\bigg|_{\gamma=0}\frac{\gamma}{t}. \tag{K.50}$$

Applying the expansion to (K.48), we get

$$E_0(t) \geq t \min_{Q_1\in\mathcal{P}(\mathcal{X})} \left(C(P_0\|Q_1) - \frac{n}{t}D(P_0\|Q_1) - (4\left|\mathcal{X}\right| + 4)\frac{\log(t+1)}{t}\right)^{+} + \alpha D(Q_1\|P_1), \tag{K.51}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where $C(P_0\|Q_1) = \max_{0\le\lambda\le 1} -\log\sum_{x\in\mathscr{X}} P_0^{1-\lambda}(x)Q_1^{\lambda}(x)$ is the Chernoff information, we have used that $\frac{\partial E(\gamma)}{\partial\gamma} = \lambda^*$ with $\lambda^* \le 1$ being the Lagrange multiplier solution of (K.48), and we have lower bounded $E_0(t)$ by setting $\lambda^* = 1$. Further lower bounding $E_0(t)$, we get

$$E_0(t) \ge t \min_{Q_1\in\mathscr{P}(\mathscr{X})} \Big(C(P_0\|Q_1) - \frac{n}{t}D(P_0\|Q_1)\Big)^+ \mathbb{1}\{t\ge\zeta n\} - (4|\mathscr{X}|+4)\frac{\log(t+1)}{t} + \alpha D(Q_1\|P_1) \tag{K.52}$$

$$\ge -(4|\mathscr{X}|+4)\log(t+1) + t \min_{Q_1\in\mathscr{P}(\mathscr{X})} \Big(C(P_0\|Q_1) - \frac{1}{\zeta}D(P_0\|Q_1)\Big)^+ \mathbb{1}\{t\ge\zeta n\} + \alpha D(Q_1\|P_1), \tag{K.53}$$

where $(x)^+ = \max\{x,0\}$. Choosing $\zeta > \frac{D(P_0\|P_1)}{C(P_0\|P_1)}$, for every $t \ge \zeta n$, the solution to the optimization (K.53) is non-zero since $\alpha D(Q_1\|P_1)$ can be zero if and only if $Q_1 = P_1$, while the first term $\Big(C(P_0\|Q_1) - \frac{1}{\zeta}D(P_0\|Q_1)\Big)^+ \mathbb{1}\{t\ge\zeta n\}$ is non-zero for that choice of Q_1. Therefore, we have

$$E_0(t) \ge -(4|\mathscr{X}|+4)\log(t+1) + E_0 t\mathbb{1}\{t\ge\zeta n\} \tag{K.54}$$

$$\ge -(4|\mathscr{X}|+4)\log(t+1) + E_0 t\Big(1-\zeta\frac{n}{t}\Big) \tag{K.55}$$

$$= -(4|\mathscr{X}|+4)\log(t+1) + E_0 t - \xi_0 n, \tag{K.56}$$

where $\xi_0 = E_0\zeta, E_0 > 0$. Substituting (K.56) in (K.44) gives (K.40). In order to prove the result under the hypothesis P_1, let

$$E_1(t) = \min_{(Q,Q_1)\in\mathscr{Q}_{11}(t)} t\big(D(Q\|P_1) + \alpha D(Q_1\|P_1)\big), \tag{K.57}$$

$$\mathscr{Q}_{11}(t) = \Big\{(Q,Q_1) : D(Q\|Q_1) - D(Q\|P_0) \ge -\frac{\gamma_{1,n}(t)}{t}\Big\}. \tag{K.58}$$

Similarly to the steps for hypothesis P_0, we have

$$E_1(t) + \lambda^*(4|\mathscr{X}|+4)\log(t+1) \tag{K.59}$$

$$\ge t \min_{Q_1\in\mathscr{P}(\mathscr{X})} \left(\max_{0\le\lambda} -\log\sum_{x\in\mathscr{X}} P_1(x)Q_1(x)^{-\lambda}P_0^{\lambda}\right) - \frac{n}{t}\lambda^* D(Q_1\|P_0) + \alpha D(Q_1\|P_1) \tag{K.60}$$

$$\ge t \min_{Q,Q_1:D(Q\|Q_1)\ge D(Q\|P_0)} D(Q\|P_1) + \frac{\alpha}{2}D(Q_1\|P_1) + \min_{Q_1}\frac{\alpha}{2}tD(Q_1\|P_1) - n\lambda^* D(Q_1\|P_0) \tag{K.61}$$

$$\ge t \min_{Q,Q_1:D(Q\|Q_1)\ge D(Q\|P_0)} D(Q\|P_1) + \frac{\alpha}{2}D(Q_1\|P_1) + n\min_{Q_1}\frac{\alpha}{2}D(Q_1\|P_1) - \lambda^* D(Q_1\|P_0) \tag{K.62}$$

$$= E_1 t - \xi_1 n, \tag{K.63}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

where in the last step we used the fact that both optimizations are finite and $E_1 > 0, \xi_1 > 0$, and where λ^* is the Lagrange multiplier solving the maximization in (K.60). This concludes the proof. □

LEMMA K.2 For $i \in \{0, 1\}$

$$\left|\gamma_{i,n}(t+1) - \gamma_{i,n}(t)\right| \xrightarrow{a.s.} 0, \tag{K.64}$$

as $t \to \infty$.

Proof. We first show the lemma for $i = 0$. By the triangle inequality and the definition of the relative entropy, we have

$$\left|\gamma_{0,n}(t+1) - \gamma_{0,n}(t)\right| \leq (4\left|\mathscr{X}\right| + 4)\left|\log(t+1) - \log(t)\right| + n\left|\sum_{x\in\mathscr{X}} P_0(x) \log \frac{\hat{T}_z'^{1:\alpha(t+1)}(x)}{\hat{T}_z'^{1:\alpha t}(x)}\right| \tag{K.65}$$

$$\leq \frac{c}{t} + n\left|\sum_{x\in\mathscr{X}} P_0(x) \log \frac{\frac{t}{t+1}\hat{T}_z'^{1:\alpha t}(x) + \frac{1}{t+1}\hat{T}_z^{\alpha t+1:\alpha(t+1)}(x)}{\hat{T}_z'^{1:\alpha t}(x)}\right| \tag{K.66}$$

$$\leq \frac{c}{t} + n\left|\sum_{x\in\mathscr{X}} P_0(x) \log \frac{t}{t+1} + \frac{1}{t+1}\frac{\hat{T}_z^{\alpha t+1:\alpha(t+1)}(x)}{\hat{T}_z'^{1:\alpha t}(x)}\right| \to 0, \tag{K.67}$$

where $\hat{T}_z^{i:j}$ is the type of the sequence $(X_i, \ldots, X_j)$, and in the last step, the logarithm will be either zero or tends to zero as t tends to infinity. Next, for $\gamma_{1,n}(t)$, by the triangle inequality and the L_1 bound on entropy [11], we have

$$\begin{aligned}\left|\gamma_{1,n}(t+1) - \gamma_{1,n}(t)\right| \leq & (4\left|\mathscr{X}\right| + 4)\left|\log(t+1) - \log(t)\right| + n\left|H(\hat{T}_z'^{1:\alpha(t+1)}) - H(\hat{T}_z'^{1:\alpha t})\right| \\ & + n\left|\sum_{x\in\mathscr{X}} (\hat{T}_z'^{1:\alpha(t+1)}(x) - \hat{T}_z'^{1:\alpha t}(x)) \log P_0(x)\right| \end{aligned}\tag{K.68}$$

$$\begin{aligned}\leq & \frac{c}{t} - n\|\hat{T}_z'^{1:\alpha(t+1)} - \hat{T}_z'^{1:\alpha t}\|_1 \log \frac{\|\hat{T}_z'^{1:\alpha(t+1)} - \hat{T}_z'^{1:\alpha t}\|_1}{\left|\mathscr{X}\right|} \\ & + c'\|\hat{T}_z'^{1:\alpha(t+1)} - \hat{T}_z'^{1:\alpha t}\|_1 \end{aligned}\tag{K.69}$$

$$\leq \frac{c}{t} + \frac{n}{t+1}\log\frac{1}{\left|\mathscr{X}\right|(t+1)} + \frac{c'}{t+1} \to 0, \tag{K.70}$$

where in the last step we have used

$$\|\hat{T}_z'^{1:\alpha(t+1)} - \hat{T}_z'^{1:\alpha t}\|_1 \leq \frac{1}{t+1}\|\hat{T}_z'^{1:\alpha t} - \hat{T}_z'^{\alpha t:\alpha(t+1)}\|_1 \leq \frac{1}{t+1}, \tag{K.71}$$

as $\hat{T}_{z'}$ is a type of a training sequence, and c, c' are positive constants. □

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Now by the finiteness of τ_0 for every n, and the definition of τ_0, there exists a finite τ_0 with probability one such that

$$S_{\tau_0-1} < \gamma_{0,n}(\tau_0 - 1), \;\; \gamma_{0,n}(\tau_0) \leq S_{\tau_0} \;\text{ w.p.1.} \tag{K.72}$$

Furthermore, for every τ_0, we have

$$\frac{S_{\tau_0}}{\tau_0} = D(\hat{T}_{\boldsymbol{x}}^{\tau_0}\|\hat{T}_{\boldsymbol{z}}^{\tau_0}) - D(\hat{T}_{\boldsymbol{x}}^{\tau_0}\|P_1) + o(1). \tag{K.73}$$

Also, since by design $\tau_0 \geq n$, using the WLLN, and the continuous mapping theorem, as $n \to \infty$, we get

$$\frac{S_{\tau_0}}{\tau_0} \xrightarrow{p} D(P_0\|P_1), \qquad \frac{S_{\tau_0-1}}{\tau_0 - 1} \xrightarrow{p} D(P_0\|P_1). \tag{K.74}$$

Therefore, by Lemma K.1, Lemma K.2, and (K.72), (K.74), we can conclude that

$$\frac{\gamma_{0,n}(\tau_0)}{\tau_0} \xrightarrow{p} D(P_0\|P_1), \tag{K.75}$$

as $n \to \infty$. Also, we have

$$\frac{\gamma_{0,n}(\tau_0)}{\tau_0} = \frac{n}{\tau_0} D(P_0\|\hat{T}'_{\boldsymbol{z}}) + \frac{\log(\tau_0 + 1)}{\tau_0}, \tag{K.76}$$

and by assumtion $\max_{x\in\mathscr{X}} \frac{P_0(x)}{P_1(x)} \leq c$, $D(P_0\|\hat{T}_{\boldsymbol{z}}{}')$ is a consistent estimator of $D(P_0\|P_1)$ [6], and hence,

$$D(P_0\|\hat{T}'_{\boldsymbol{z}}) \xrightarrow{p} D(P_0\|P_1). \tag{K.77}$$

Finally by (K.75), (K.76), (K.77) and using continuous mapping theorem [33], we have

$$\frac{\tau_0}{n} \xrightarrow{p} 1, \tag{K.78}$$

as $n \to \infty$.

To show the convergence in L^1, we only need to prove the uniform integrability of the sequence of random variables $\frac{\tau_0}{n}$, where τ_0 depends on n. Equivalently, we need to show that

$$\lim_{t\to\infty} \sup_{n\geq 1} \mathbb{E}_{P_0}\left[\frac{\tau_0}{n}\mathbb{1}\left\{\frac{\tau_0}{n} \geq t\right\}\right] = 0. \tag{K.79}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

By (K.40), we can upper bound the given expectation in (K.79) as

$$\mathbb{E}_{P_0}\left[\frac{\tau_0}{n}\mathbb{1}\{\tau_0 \geq tn\}\right] = \frac{1}{n}\sum_{m=1}^{\infty}\mathbb{P}_0[\tau_0 - tn \geq m] \tag{K.80}$$

$$\leq \frac{1}{n}te^{-n(tE_0-\xi_0)}\sum_{m=0}^{\infty}c(m+tn+1)^{4|\mathscr{X}|}e^{-mE_0}. \tag{K.81}$$

Hence the expectation is vanishing as $t \to \infty$ for every n giving the uniform integrability of $\frac{\tau_0}{n}$, and hence convergence in L^1 [2], i.e.

$$\lim_{n\to\infty}\mathbb{E}_{P_0}\left[\left|\frac{\tau_0}{n}-1\right|\right] = 0. \tag{K.82}$$

Finally, we prove the convergence of τ. By (K.38), (K.75) and the union bound, we obtain

$$\mathbb{P}_0\left[\left|\frac{\tau}{n}-1\right| \geq \varepsilon\right] \leq \mathbb{P}_0\left[\left|\frac{\tau}{n}-1\right| \geq \varepsilon, \phi = 0\right] + \mathbb{P}_0[\phi = 1] \tag{K.83}$$

$$= \mathbb{P}_0\left[\left|\frac{\tau_0}{n}-1\right| \geq \varepsilon\right] + \varepsilon_0, \tag{K.84}$$

which tends to 0 as $n \to \infty$, establishing the convergence of $\frac{\tau}{n}$ in probability. Now, using that $\tau \leq \tau_0$, we have

$$\mathbb{E}_{P_0}\left[\frac{\tau}{n}\mathbb{1}\left\{\frac{\tau}{n} \geq t\right\}\right] \leq \mathbb{E}_{P_0}\left[\frac{\tau_0}{n}\mathbb{1}\left\{\frac{\tau_0}{n} \geq t\right\}\right]. \tag{K.85}$$

Therefore, uniform integrability of τ_0 gives the uniform integrability of τ, and hence convergence in L^1 norm and also expectation of $\frac{\tau}{n}$, which concludes the proof.

L. Proof of Theorem 5.3

For the type-II error exponent, the converse for sequential hypothesis testing is applicable, i.e. for every sequential test with $\mathbb{E}_{P_0}[\tau] \leq n$, we have $E_1 \leq D(P_0\|P_1)$ [32]. In order to find an upper bound to E_0, we use the following lemma.

Lemma L.1 For any type-based sequential test Φ^{seq}, let $\tau \in \mathscr{N}_1^{\varepsilon}$, where $\mathscr{N}_1^{\varepsilon} = \{1, ..., t\}$ is the typical stopping time set such that

$$P_1(\tau \in \mathscr{N}_1^{\varepsilon}) \geq 1 - \frac{\varepsilon}{2}. \tag{L.1}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

Also for every t, let $(\boldsymbol{x}^t, \boldsymbol{z}^{\alpha t}) \in \mathscr{B}_t^{\varepsilon}$ where

$$\mathscr{B}_t^{\varepsilon} = \left\{ (\boldsymbol{x}^t, \boldsymbol{z}^{\alpha t}) : \max_{a \in \mathscr{X}} \left\{ \left| \hat{T}_{\boldsymbol{x}}(a) - \hat{T}_{z}(a) \right| \right\} \leq \sqrt{\frac{2}{(\alpha \wedge 1)t} \log \frac{8|\mathscr{X}|}{\varepsilon}} \right\}. \tag{L.2}$$

Then for any type-based test $(\phi(\hat{T}_{\boldsymbol{x}}, \hat{T}_z), \tau)$ such that for all distributions $P_1 \in \mathscr{P}(\mathscr{X})$,

$$\varepsilon_1\big(\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau})\big) < \varepsilon, \quad \varepsilon \in \Big(0, \frac{1}{2}\Big), \tag{L.3}$$

we have

$$\mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) = 1, (\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon}, \tau \in \mathscr{N}_1^{\varepsilon}\Big] \geq 1 - 2\varepsilon. \tag{L.4}$$

Proof. We prove the lemma by contradiction. Assume

$$\mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) = 1, (\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon}, \tau \in \mathscr{N}_1^{\varepsilon}\Big] < 1 - 2\varepsilon. \tag{L.5}$$

Then,

$$2\varepsilon < \mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \neq 1 \cup (\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \notin \mathscr{B}_{\tau}^{\varepsilon} \cup \tau \notin \mathscr{N}_1^{\varepsilon}\Big] \tag{L.6}$$

$$\leq \mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \neq 1\Big] + \mathbb{P}\Big[(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \notin \mathscr{B}_{\tau}^{\varepsilon}\Big] + \mathbb{P}\Big[\tau \notin \mathscr{N}_1^{\varepsilon}\Big] \tag{L.7}$$

$$\leq \varepsilon_1 + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2}, \tag{L.8}$$

where in the last step we used Lemma J.1. Hence, $\varepsilon \leq \varepsilon_1$ which is a contradiction and hence (L.4) holds. □

By Lemma L.1, we can conclude that if condition (L.4) does not hold, there exists a distribution P_1 such that the type-II error probability is bounded away from zero and hence the type-II error exponent of such test equals to zero. Therefore, by (L.4), we can lower bound the ε_0 for any test with non-zero E_1 as

$$\varepsilon_0 = \mathbb{P}[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) = 1] \tag{L.9}$$

$$\geq \mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) = 1, (\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon}, \tau \in \mathscr{N}_1^{\varepsilon}\Big] \tag{L.10}$$

$$= \mathbb{P}\Big[\phi(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) = 1 \Big| (\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon}, \tau \in \mathscr{N}_1\Big] \mathbb{P}\Big[(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon}, \tau \in \mathscr{N}_1\Big]. \tag{L.11}$$

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024

By the previous lemma, we can lower bound the first probability by $1-2\varepsilon$. Also, let ε to be sufficiently small, such that $n \in \mathscr{N}_1^{\varepsilon} = \{1, ..., N\}$, then by the method of types

$$\varepsilon_0 \geq (1-2\varepsilon)\sum_{t=1}^{N} \mathbb{P}\Big[(\boldsymbol{x}^{\tau}, \boldsymbol{z}^{\alpha\tau}) \in \mathscr{B}_{\tau}^{\varepsilon} \Big| \tau = t\Big]\mathbb{P}_0[\tau = t] \tag{L.12}$$

$$\geq (1-2\varepsilon)\sum_{t=1}^{n} (t+1)^{-\left|\mathscr{X}\right|}(\alpha t+1)^{-\left|\mathscr{X}\right|} e^{-t \min_{(Q,Q_1)\in\mathscr{B}_t^{\varepsilon}} D(Q\|P_0)+\alpha D(Q_1\|P_1)}\mathbb{P}_0[\tau = t] \tag{L.13}$$

$$\geq (1-2\varepsilon)ce^{-n \min_{(Q,Q_1)\in\mathscr{B}_t^{\varepsilon}} D(Q\|P_0)+\alpha D(Q_1\|P_1)}\mathbb{P}_0[\tau \leq n], \tag{L.14}$$

where c is a positive constant. Now by $\mathbb{E}_{P_0}[\tau] \leq n$, the optimal test should stop by time n with positive probability, i.e. $\mathbb{P}_0[\tau \leq n] > 0$, since otherwise, $\mathbb{E}_{P_0}[\tau] > n$. Finally, by letting $\varepsilon \to 0, n \to \infty$, we have

$$E_0 \leq D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0). \tag{L.15}$$

Hence, for any sequential test with a finite stopping time, and the type-II error probability that is bounded away from one for every distribution P_1, the type-I error exponent is bounded by $D_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$, which concludes the proof.

Downloaded from https://academic.oup.com/imaiai/article/13/2/iaae010/7664051 by Hutchison/MRC Research Centre user on 20 June 2024