A Competitive Neyman-Pearson Approach to Universal Hypothesis Testing with Applications *

Evgeny Levitan and Neri Merhav †

December 4, 2000

Abstract – The problem of hypothesis testing for parametric information sources whose parameters are not explicitly known is considered. A new, modified version of the Neyman-Pearson criterion of optimality, where the uniform constraint on exponential rate of the false-alarm probability is replaced by a one that depends on unknown values of the parameters, is proposed. An optimal universal decision rule, based on Kullback-Leibler divergence, is developed and shown to be efficient in the sense of achieving exponential decay of both mis-detection and false-alarm probabilities for *all* values of unknown parameters, whenever such an efficient decision rule at all exists. Furthermore, necessary and sufficient conditions for the existence of such efficient universal tests are established and the best universally achievable error exponents are presented. Finally, the proposed approach is applied to several important problems in signal processing and communications and compared to the generalized likelihood ratio test.

Index Terms – Hypothesis testing, universal tests, Neyman-Pearson, likelihood ratio test, generalized likelihood ratio test, error exponent, universal classification, model order estimation, universal decoding.

---

*This research is supported by the Israeli Science Foundation.

†The authors are with the Electrical Engineering Department, Technion - I.I.T. Haifa 32000, Israel. E-mail: levitan@techunix.technion.ac.il, merhav@ee.technion.ac.il.

1 Introduction

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be a sequence of random variables which take on values in a certain finite alphabet A. The binary hypothesis testing problem is that of deciding, based on observing $\boldsymbol{y}$, whether this sequence has originated from a source with a probability distribution P_{θ_1} (hypothesis H_1) or from a source with a probability distribution P_{θ_2} (hypothesis H_2). The distribution P_{θ_i}, associated with the hypothesis H_i, is known to belong to a certain parametric family of probability mass functions (PMF's) $\{p_{\theta_i}(\boldsymbol{y}), \theta_i \in \Theta_i\}$, where θ_i is the parameter of the PMF within the family and Θ_i is the parameter set, $i = 1, 2$.

A decision rule Ω is a sequence of partitions $\Omega^n = (\Omega_1^n, \Omega_2^n)$ $(n = 1, 2, \ldots)$ of the observation space A^n into two complementary regions Ω_1^n and Ω_2^n whose union equals A^n, with the interpretation that for $\boldsymbol{y} \in \Omega_i^n$, a decision is made in favor of hypothesis H_i, $i = 1, 2$.

Let $P_{e_1}(\Omega^n|\theta_1) \triangleq p_{\theta_1}(\boldsymbol{y} \in \Omega_2^n)$ and $P_{e_2}(\Omega^n|\theta_2) \triangleq p_{\theta_2}(\boldsymbol{y} \in \Omega_1^n)$ denote the first (false-alarm) and the second (mis-detection) kinds of error probability, respectively. The classical Neyman-Pearson approach [1] to simple binary hypothesis testing (θ_1 and θ_2 are known) suggests to minimize the probability of error of the second kind $P_{e_2}(\Omega^n|\theta_2)$ subject to the constraint that the probability of error of the first kind $P_{e_1}(\Omega^n|\theta_1)$ is less than $2^{-\lambda n}$ for some $\lambda > 0$. An alternative approach of interest is the Bayes criterion, in which a decision rule is sought to minimize the overall probability of error given by $P_e(\Omega^n|\theta_1, \theta_2) \triangleq \pi_1 P_{e_1}(\Omega^n|\theta_1) + \pi_2 P_{e_2}(\Omega^n|\theta_2)$, where π_1 and π_2 are prior probabilities of the hypotheses. The optimal test under both criteria is well-known [2] to be the likelihood ratio test (LRT), which compares the likelihood ratio $p_{\theta_2}(\boldsymbol{y})/p_{\theta_1}(\boldsymbol{y})$ to a suitable threshold in order to make a decision.

In many problems of practical importance the situation is not so simple and θ_1 and θ_2 are not fully known. All one knows is that the parameters θ_1 and θ_2 take on values in two disjoint sets Θ_1 and Θ_2, respectively. In this case, H_1

and H_2 are referred to as composite hypotheses. If θ_i is treated as a random variable with known prior probability density (Bayesian approach), then composite hypothesis testing problem can be reduced to the simple one by averaging $p_{\theta_i}(\boldsymbol{y})$ over θ_i. Therefore, the LRT with respect to these mixture densities can be implemented to minimize the average (over θ_1 and θ_2) probability of error.

In many cases of interest, however, it is unrealistic to consider the unknown parameter as a random variable and it is assumed to be fixed. Since θ_i is unknown and has no probability law, the LRT can not be applied and hence the aim is to design another test, which is *universal* in the sense that it does not depend on θ_1 and θ_2 and, nonetheless, performs well in a certain sense for every θ_1 and θ_2. In this situation, a generalized notion of the Neyman-Pearson criterion, originally proposed in [3], is frequently used (see e.g., [4] - [11]). According to this criterion, an optimal decision rule is sought to maximize the mis-detection exponent uniformly over all possible probability laws P_{θ_2}, subject to the constraint that for every P_{θ_1} the false-alarm exponent is not less than a given $\lambda > 0$. Mathematically, the criterion is:

$$\sup_{\Omega} \liminf_{n\to\infty} -\frac{1}{n} \log P_{e_2}(\Omega^n|\theta_2), \quad \forall\theta_2 \in \Theta_2 \tag{1}$$

s.t.

$$\liminf_{n\to\infty} -\frac{1}{n} \log P_{e_1}(\Omega^n|\theta_1) \geq \lambda, \quad \forall\theta_1 \in \Theta_1. \tag{2}$$

An optimal test under this criterion was first developed by Hoeffding [3] for i.i.d. sources over a finite alphabet, when the hypothesis H_1 is simple and H_2 is composite. It was later generalized to Markov models, continuous alphabet and composite hypotheses [12], [13]. The generalized version of the Hoeffding test is based on comparing the worst case of relative entropy (informational divergence) between the empirical measure associated with $\boldsymbol{y}$ and the probability distribution under H_1 to the threshold λ, i.e.,

$$\Omega_1^n = \left\{ \boldsymbol{y} : \inf_{\theta_1\in\Theta_1} D(Q_{\boldsymbol{y}}||P_{\theta_1}) < \lambda \right\}. \tag{3}$$

Also, in several composite hypothesis testing problems considered in [4]-[9], another test, known as the generalized likelihood ratio test (GLRT), was shown to be asymptotically optimal under the generalized Neyman-Pearson criterion. The GLRT, which sometimes coincides with (3), uses maximum likelihood (ML) estimates of θ_1 and θ_2 under H_1 and H_2, respectively, to implement an LRT. In other words, this test compares the generalized likelihood ratio $\sup_{\theta_2 \in \Theta_2} p_{\theta_2}(\boldsymbol{y}) / \sup_{\theta_1 \in \Theta_1} p_{\theta_1}(\boldsymbol{y})$ to a certain threshold. In some situations, the GLRT is asymptotically optimal also in the Bayesian sense [14], minimax sense [15] and random coding sense [16]. Although the GLRT is not always optimal [17], [18, Appendix], it is widely used in universal hypothesis testing because in most of the practical situations this approach gives satisfactory results.

While universal decision rules are independent of the unknown parameters θ_1 and θ_2, the performance, in general, will depend on them. We are usually interested in exponential decay of the error probabilities, and we say that a universal test is *efficient*, if it achieves exponential decay of both error probabilities for all values of θ_1 and θ_2. Thus, the important objective in the generalized Neyman-Pearson approach is that of choosing the threshold λ such that the second kind error probability will vanish exponentially fast with n for every θ_2. As shown in [4]-[9], for every distinct P_{θ_1} and P_{θ_2}, there exists some $\lambda > 0$ such that the error probabilities under both hypotheses decay exponentially to zero. However, such λ depends on the true underlying probability measures that are in turn unknown. Therefore, to assure exponential decay of both error probabilities for all P_{θ_1} and P_{θ_2}, the value of λ should be selected small enough. Moreover, if the families of PMF's associated with the hypotheses are sufficiently rich, such $\lambda > 0$ does not exist at all. It means that for any λ, one can find P_{θ_1} and P_{θ_2}, which are close enough each to other, so that the requirement (2) will be too restrictive, and even an optimal LRT that satisfies (2) for these θ_1 and θ_2 will not be able to discriminate between the hypotheses, i.e., the mis-detection probability will tend to unity [4, Theorem 3], [8, Remark 1(b)].

One customary approach to overcome this difficulty is to adjust the threshold λ empirically during an experiment, as was also suggested in [8]. But in this case, the universality property of a test will be broken off, since one can not specify a good value of λ before an experiment. Another method, proposed in [19], [20], is to let λ tend to zero with sufficiently small rate as $n \to \infty$. This approach, unfortunately, achieves exponentially vanishing mis-detection probability at the cost of the exponential decay of the false-alarm probability, that is $P_{e_1}(\Omega^n|\theta_1)$ will decay only subexponentially to zero.

In this paper, we propose a new, *competitive* version of the Neyman-Pearson criterion to composite hypothesis testing, that could potentially solve the inconsistency problem described above. The main idea behind this approach is to replace the uniform constraint on the error rate under H_1 by a softer one. Specifically, we wish to find a decision rule that maximizes the second kind error exponent uniformly over θ_2, subject to the following condition:

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_{e_1}(\Omega^n|\theta_1) \geq \lambda(\theta_1,\theta_2), \quad \forall P_{\theta_1}, P_{\theta_2}, \tag{4}$$

where $\lambda(\theta_1,\theta_2)$ is a certain threshold function that determines, for every θ_1 and θ_2, the minimal allowable exponential rate of the first kind error probability. Observe that this is just a modified version of the generalized Neyman-Pearson approach where λ is allowed to be a function of the parameters θ_1 and θ_2. This modification adapts the constraint on the first kind error rate to the ability of discrimination between P_{θ_1} and P_{θ_2}, as measured by $\lambda(\theta_1,\theta_2)$. Namely, for hardly distinguishable hypotheses, $\lambda(\theta_1,\theta_2)$ is expected to take on small values and hence the constraint (4) turns to be weaker than (2), whereas for those parameters θ_1 and θ_2 that $\lambda(\theta_1,\theta_2)$ takes on relatively large values, a higher error rate under H_1 is required. In the radar detection problem, for example, it would be desirable to adjust the false-alarm rate according to an unknown level of the signal-to-noise ratio (SNR).

An optimal test under this competitive Neyman-Pearson criterion is quite a straightforward extension of that under the generalized Neyman-Pearson approach that was derived in the previous works on universal hypotheses testing [3]-[13]. The major interest in the proposed approach, however, is in specifying a reasonable threshold function $\lambda(\theta_1, \theta_2)$ that would lead to an efficient universal decision rule.

In a recent paper [18], that in fact has motivated our work, a novel competitive minimax approach to composite hypothesis testing was proposed for the Bayesian setting. An optimal decision rule in the competitive minimax sense minimizes the worst-case ratio between the error probability of the test that is independent of the unknown (θ_1, θ_2) and the minimum error probability achieved by the LRT. That is,

$$K_n \triangleq \inf_{\Omega^n} \sup_{\theta_1, \theta_2} \frac{P_e(\Omega^n|\theta_1, \theta_2)}{P_e^*(\theta_1, \theta_2)}, \tag{5}$$

where $P_e(\Omega^n|\theta_1, \theta_2)$ is the overall probability of error associated with a decision rule Ω^n and $P_e^*(\theta_1, \theta_2)$ is the minimum Bayes error probability of the LRT for known θ_1 and θ_2. If K_n happens to be subexponential in n, then an optimal sequence of decision rules under the competitive minimax criterion attains the same exponential error rate as the optimum LRT for every θ_1 and θ_2. On the other hand, when K_n grows exponentially with n, this criterion, unfortunately, does not guarantee an exponential decay of the error probability, and therefore the following modification has been proposed:

$$K_n^\xi \triangleq \inf_{\Omega^n} \sup_{\theta_1, \theta_2} \frac{P_e(\Omega^n|\theta_1, \theta_2)}{[P_e^*(\theta_1, \theta_2)]^\xi}, \tag{6}$$

where $0 \le \xi \le 1$ is selected to be the largest number ξ^* such that K_n^ξ does not grow exponentially with n. An asymptotically minimax-optimal test developed in [18] is given by

$$\widehat{\Omega}_1^n = \left\{ \boldsymbol{y} : \sup_{\theta_1, \theta_2} \frac{p_{\theta_1}(\boldsymbol{y})}{[P_e^*(\theta_1, \theta_2)]^{\xi^*}} > \sup_{\theta_1, \theta_2} \frac{p_{\theta_2}(\boldsymbol{y})}{[P_e^*(\theta_1, \theta_2)]^{\xi^*}} \right\}. \tag{7}$$

Actually, this decision rule asymptotically achieves the maximal fraction ξ^* of the optimum error exponent and is therefore asymptotically equivalent, under certain regularity conditions, to the test that maximizes the worst-case ratio between the exponential error rates of a decision rule that is ignorant of (θ_1, θ_2) and the LRT, specifically,

$$\sup_{\Omega} \inf_{\theta_1, \theta_2} \frac{\liminf_{n\to\infty} -\frac{1}{n}\log P_e(\Omega^n|\theta_1, \theta_2)}{E^*(\theta_1, \theta_2)}, \tag{8}$$

where $E^*(\theta_1, \theta_2) = \liminf_{n\to\infty} -\frac{1}{n}\log P_e^*(\theta_1, \theta_2)$ is the exponential error rate associated with the LRT.

To see the interrelation between this approach and the competitive Neyman-Pearson criterion, consider a specific choice of $\lambda(\theta_1, \theta_2) = \xi E^*(\theta_1, \theta_2)$, where $\xi > 0$ is a given number. Then, condition (4) restricts consideration to tests whose worst-case value of the ratio between the first kind error exponent and the optimum error exponent of the LRT is not less than ξ. More precisely, the constraint on the first kind error probability can be rewritten in the following form:

$$\inf_{\theta_1, \theta_2} \frac{\liminf_{n\to\infty} -\frac{1}{n}\log P_{e_1}(\Omega^n|\theta_1)}{E^*(\theta_1, \theta_2)} \geq \xi, \tag{9}$$

where $\xi > 0$ designates the maximal tolerable level of the loss (or the minimal gain) in the false-alarm rate relative to $E^*(\theta_1, \theta_2)$ caused by uncertainty in (θ_1, θ_2). In light of these observations, our competitive Neyman-Pearson approach may be viewed as an extension of the competitive minimax criterion to a Neyman-Pearson-like setting of the composite hypothesis testing problem.

In our work we propose and investigate a universal decision rule, which is optimal in this competitive Neyman-Pearson sense. We also derive a single-letter expression for the second kind error exponent and establish the necessary and sufficient condition on $\lambda(\theta_1, \theta_2)$ under which exponential decay of the mis-detection probability is guaranteed for all θ_2. Generally speaking, our main result is that $\inf_{\theta_2' \in \Theta_2} D(P_{\theta_2'}||P_{\theta_1})$ serves as the supremum over all error rates under H_1 that could be achieved by a universal decision rule, which still guarantees, for all $\theta_2 \in \Theta_2$, exponential decay of the second kind error probability. In effect, it can be seen as generalization of Stein's Lemma (cf. e.g., [16, Corollary

1.2]) to composite hypotheses. The significance of this result is that it enables us to establish conditions on the richness and the structure of the parameter sets under which it is possible to distinguish efficiently between the hypotheses. In addition, we develop an optimal decision rule under the modified version of the competitive minimax criterion (with ξ^*) for the Bayesian setting of the composite hypothesis testing problem. In contrast to the competitive minimax test (7), this decision rule is shown to be independent of ξ^*, which is normally unavailable in closed form. Finally, we present applications of the proposed approach to problems of classification with training sequences, model order estimation and detection of messages via unknown channels. The performance will be examined and compared to the generalized Neyman-Pearson approach and other existing methods.

For the sake of simplicity, the general analysis will be restricted to the case of i.i.d. sources with a finite alphabet, but it can easily be extended to Markov sources, finite-state (FS) arbitrary varying sources (AVSs) with known deterministic state sequences and more general alphabets. In addition, using an appropriate definition of the Neyman-Pearson criterion for multiple hypotheses testing, a generalization to M hypotheses is also possible as long as M does not grow exponentially with n. The analysis techniques that will be used are similar to those of [4].

The remainder of the paper is organized as follows. In the next section, the problem is precisely formulated and main results are derived. Section 3 contains the aforementioned applications. Finally, in Section 4, we summarize our conclusions.

2 Statement of the Problem and Main Results

2.1 Statement of the Problem

Let $\mathbf{P}_i \triangleq \{p_{\theta_i}(\cdot) : \theta_i \in \Theta_i\}$ denote a parametric family of memoryless sources with a finite alphabet A, whose cardinality is $|A|$, where θ_i is a parameter vector consisting of the *strictly positive* letter probabilities, and $\Theta_i \subseteq \mathbb{R}^{|A|-1}$ is a parameter set, $i = 1, 2$. Let $P_{\theta_1} \in \mathbf{P}_1$ and $P_{\theta_2} \in \mathbf{P}_2$ be two sources in these families. The assumption about positivity of the letter probabilities is needed to guarantee that the functional $D(\cdot||P_{\theta_i})$, which is defined below in (15), is continuous. This is required for the proof of Theorem 2 in Subsection 2.2. Let $\theta \triangleq (\theta_1, \theta_2) \in \Theta$, where Θ denotes the Cartesian product $\Theta_1 \times \Theta_2$. The unknown θ is assumed fixed and deterministic. Given a sequence of observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in A^n$, we wish to decide between two hypotheses $\{H_i, i = 1, 2\}$, where under H_i it is assumed that $\boldsymbol{y}$ was emitted from the source P_{θ_i}.

The probability of error under hypothesis H_i, associated with a decision rule $\Omega^n = (\Omega_1^n, \Omega_2^n)$, is given by

$$P_{e_i}(\Omega^n|\theta_i) = \sum_{\boldsymbol{y} \in (\Omega_i^n)^c} p_{\theta_i}(\boldsymbol{y}), \quad i = 1, 2, \tag{10}$$

where $(\Omega_i^n)^c$ is the complement set of Ω_i^n and $p_{\theta_i}(\boldsymbol{y})$ is the conditional PMF of $\boldsymbol{y}$ given θ_i. Let $e_1(\Omega|\theta_1)$ and $e_2(\Omega|\theta_2)$ denote the first and the second kind error exponents, respectively, associated with a sequence Ω of decision rules Ω^n ($n = 1, 2 \ldots$) and induced by θ, i.e.,

$$e_i(\Omega|\theta_i) \triangleq \liminf_{n\to\infty} -\frac{1}{n} \log P_{e_i}(\Omega^n|\theta_i), \quad i = 1, 2. \tag{11}$$

We wish to find an optimal decision rule in the competitive Neyman-Pearson sense, that is, among all tests satisfying

$$e_1(\Omega|\theta_1) \geq \lambda(\theta), \quad \forall\theta \in \Theta, \tag{12}$$

the optimal test will maximize $e_2(\Omega|\theta_2)$ uniformly over $\theta_2 \in \Theta_2$, where $\lambda(\theta)$ is an arbitrary nonnegative threshold function. In other words, we seek a decision rule, that for every P_{θ_1} and P_{θ_2} achieves exponential decay of the false-alarm probability with rate at least $\lambda(\theta)$, and at the same time maximizes the mis-detection exponent, whatever the true underlying probability measures are. Our goal is to analyze the performance of this optimal decision rule and to establish conditions on the threshold function under which both error probabilities vanish exponentially fast with n for all P_{θ_1} and P_{θ_2}.

2.2 Main Results

Let $q_{\boldsymbol{y}}(\alpha)$ denote the relative frequency of appearance of the letter $\alpha \in A$ in the vector $\boldsymbol{y} \in A^n$

$$q_{\boldsymbol{y}}(\alpha) = \frac{1}{n}\sum_{j=1}^{n} \delta(y_j = \alpha), \tag{13}$$

where $\delta(y_j = \alpha)$ is an indicator function for $y_j = \alpha$. Since $Q_{\boldsymbol{y}} \triangleq \{q_{\boldsymbol{y}}(\alpha) : \alpha \in A\}$ is a probability measure over the finite alphabet A, we define the empirical entropy and the divergence

$$H(Q_{\boldsymbol{y}}) = -\sum_{\alpha \in A} q_{\boldsymbol{y}}(\alpha) \log q_{\boldsymbol{y}}(\alpha), \tag{14}$$

$$D(Q_{\boldsymbol{y}}||P_{\theta_i}) = \sum_{\alpha \in A} q_{\boldsymbol{y}}(\alpha) \log \frac{q_{\boldsymbol{y}}(\alpha)}{p_{\theta_i}(\alpha)}, \tag{15}$$

where logarithms here and throughout the sequel are taken to the base 2 and $0 \log 0 \triangleq 0$. Note that, as was mentioned above, $D(\cdot||P_{\theta_i})$ is continuous since $p_{\theta_i}(\alpha)$ is assumed positive for every $\alpha \in A$. The type class $T(Q_{\boldsymbol{y}})$ is defined as the set of all sequences $y' \in A^n$ for which $Q_{\boldsymbol{y'}} = Q_{\boldsymbol{y}}$. It is well-known [16] that

$$p_{\theta_i}(\boldsymbol{y}) = \exp_2\left\{-n[H(Q_{\boldsymbol{y}}) + D(Q_{\boldsymbol{y}}||P_{\theta_i})]\right\}. \tag{16}$$

Let a decision rule Λ be defined as

$$\Lambda_1^n = \left\{\boldsymbol{y} : \inf_{\theta \in \Theta} \Big(D(Q_{\boldsymbol{y}}||P_{\theta_1}) - \lambda(\theta)\Big) < 0\right\}. \tag{17}$$

In the following theorem we state that Λ is an asymptotically optimal test in the competitive Neyman-Pearson sense.

Theorem 1. *Let the decision rule* Λ *be defined as in* (17).

(a) For every $\theta \in \Theta$

$$e_1(\Lambda|\theta_1) \geq \lambda(\theta). \tag{18}$$

(b) Let Ω *be an arbitrary sequence of partitions* $\Omega^n = (\Omega_1^n, \Omega_2^n)$ $(n = 1, 2, \ldots)$ *based on* $\boldsymbol{y}$*, which is independent of the true underlying probability measures, and that at the same time satisfies*

$$-\frac{1}{n} \log P_{e_1}(\Omega^n|\theta_1) \geq \lambda(\theta) + \rho_n, \quad \forall \theta \in \Theta, \tag{19}$$

where for all n *sufficiently large* $\rho_n \geq |A| \log(n+1)/n$. *Then:*

$$e_2(\Lambda|\theta_2) \geq e_2(\Omega|\theta_2), \quad \forall \theta_2 \in \Theta_2. \tag{20}$$

The optimality of Λ, in the sense of Theorem 1, essentially means that if the guaranteed performance, in terms of the first kind error probability, of an arbitrary competing decision rule Ω is slightly better than that of Λ, then Ω is inferior to Λ in the second kind error exponent uniformly for every P_{θ_2}.

Proof of Theorem 1. Since P_{θ_1} and P_{θ_2} are memoryless sources, it can be proved that $Q_{\boldsymbol{y}}$ is a sufficient statistic for asymptotic optimality. Namely, for every decision rule Ω there exists another decision rule, based only on the empirical statistic $Q_{\boldsymbol{y}}$ of the observed data $\boldsymbol{y}$, which is not worse than Ω in the error exponents sense (see, e.g., [13, Lemma 1]). Hence, we may restrict ourselves to those tests which depend on $\boldsymbol{y}$ only via $Q_{\boldsymbol{y}}$, without loss of generality. Thus,

$$P_{e_1}(\Omega^n|\theta_1) = \sum_{\boldsymbol{y} \in \Omega_2^n} p_{\theta_1}(\boldsymbol{y}) \tag{21}$$

$$= \sum_{T(Q_{\boldsymbol{y}}) \subseteq \Omega_2^n} |T(Q_{\boldsymbol{y}})| \cdot p_{\theta_1}(\boldsymbol{y}), \tag{22}$$

where $|T(Q_{\boldsymbol{y}})|$ is the size of the type class $T(Q_{\boldsymbol{y}})$. The cardinality of the type class is well-known [16] to be bounded as follows:

$$\exp_2\{n[H(Q_{\boldsymbol{y}}) - \epsilon_n]\} \leq |T(Q_{\boldsymbol{y}})| \leq \exp_2\{nH(Q_{\boldsymbol{y}})\}, \tag{23}$$

where $\epsilon_n = |A|\log(n+1)/n$. Combining this with (16), and using the constraint (19) on the first kind error probability, we have that for any $\boldsymbol{y} \in \Omega_2^n$ and $\theta \in \Theta$

$$2^{-n[\lambda(\theta)+\rho_n]} \geq P_{e_1}(\Omega^n|\theta_1) \tag{24}$$

$$\geq \sum_{T(Q_{\boldsymbol{y}})\subseteq\Omega_2^n} \exp_2\{-n[D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \epsilon_n]\} \tag{25}$$

$$\exp_2\{-n[D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \epsilon_n]\}. \tag{26}$$

Since $\rho_n \geq \epsilon_n$, we conclude that for sufficiently large n and all $\boldsymbol{y} \in \Omega_2^n$

$$D(Q_{\boldsymbol{y}}||P_{\theta_1}) \geq \lambda(\theta), \quad \forall\theta \in \Theta, \tag{27}$$

and therefore,

$$\inf_{\theta\in\Theta} \Big(D(Q_{\boldsymbol{y}}||P_{\theta_1}) - \lambda(\theta)\Big) \geq 0, \quad \forall \boldsymbol{y} \in \Omega_2^n. \tag{28}$$

It means that for n sufficiently large, any $\boldsymbol{y}$ that belongs to Ω_2^n is also in Λ_2^n, or equivalently, $\Lambda_1^n \subseteq \Omega_1^n$. Hence,

$$P_{e_2}(\Lambda^n|\theta_2) \leq P_{e_2}(\Omega^n|\theta_2), \quad \forall\theta_2 \in \Theta_2 \tag{29}$$

and part (b) is proved.

As for the part (a), we have for all $\theta \in \Theta$

$$P_{e_1}(\Lambda^n|\theta_1) = \sum_{T(Q_{\boldsymbol{y}})\subseteq\Lambda_2^n} |T(Q_{\boldsymbol{y}})| \cdot p_{\theta_1}(\boldsymbol{y}) \tag{30}$$

$$\leq \sum_{T(Q_{\boldsymbol{y}})\subseteq\Lambda_2^n} \exp_2\{-nD(Q_{\boldsymbol{y}}||P_{\theta_1})\} \tag{31}$$

$$\leq \sum_{T(Q_{\boldsymbol{y}}))\subseteq\Lambda_2^n} \exp_2\{-n\lambda(\theta)\}, \tag{32}$$

where the last inequality follows from the definition of Λ_2^n. Since the number of distinct empirical measures $Q_{\boldsymbol{y}}$ is upper-bounded by $(n+1)^{|A|}$ [16], we obtain

$$P_{e_1}(\Lambda^n|\theta_1) \leq \exp_2\{-n[\lambda(\theta) - \epsilon_n]\} \tag{33}$$

Consequently,

$$e_1(\Lambda|\theta_1) \geq \lambda(\theta), \quad \forall\theta \in \Theta, \tag{34}$$

and the theorem is proved. $\square$

Although Theorem 1 provides an optimal decision rule for any $\lambda(\theta)$, it does not directly specify an asymptotic behavior of the second kind error probability, which depends on the unknown P_{θ_2} as well as on the threshold function $\lambda(\theta)$ that can be adjusted by a detector. The fundamental question that has to be considered is whether there exists such $\lambda(\theta)$ that guarantees exponential decay of both error probabilities for all possible sources P_{θ_1} and P_{θ_2}. If so, what would be a reasonable choice of $\lambda(\theta)$?

For convenience, let us define

$$g(Q) \triangleq \inf_{\theta' \in \Theta} \Big[D(Q||P_{\theta_1'}) - \lambda(\theta') \Big], \tag{35}$$

where Q is a PMF on the alphabet A. Also, for any memoryless PMF with parameter vector μ, we denote by $B(\mu, \delta)$ an open ball of radius $\delta > 0$ around μ taken in some metric in the Euclidean space.

Throughout the sequel, for any set S, $\overline{S}$ denotes the closure of S, S^o the interior of S, and S^c the complement of S.

The next theorem determines the second kind error exponent associated with the optimal decision rule Λ and establishes a condition on $\lambda(\theta)$ under which $P_{e_2}(\Lambda^n|\theta_2)$ decays exponentially fast to zero.

Theorem 2. *Let the decision rule Λ be defined as in (17). Let P_{θ_1} and P_{θ_2} be the true underlying probability measures with unknown $\theta \in \Theta$. Then,*

a)

$$e_2(\Lambda|\theta_2) = \inf_{Q \in C} D(Q||P_{\theta_2}), \tag{36}$$

where C is the set of all PMF's over the finite alphabet A, defined as

$$C = \Big\{ Q : g(Q) < 0 \Big\}. \tag{37}$$

b)

$$e_2(\Lambda|\theta_2) > 0 \tag{38}$$

if and only if there exists some $\delta(\theta_2) > 0$ such that

$$\lambda(\theta') \leq \inf_{\mu \in B(\theta_2, \delta(\theta_2))} D(P_\mu||P_{\theta_1'}), \quad \forall \theta' \in \Theta. \tag{39}$$

Discussion: Part (a) of this theorem provides a single-letter expression for the second kind error exponent as a functional of θ_2 and $\lambda(\cdot)$. Part (b) specifies the necessary and sufficient condition on $\lambda(\cdot)$ to attain exponential decay of the second kind error probability for a specific value of θ_2. Observe, that this condition is expressed as an upper bound on $\lambda(\cdot)$ that depends on the unknown θ_2. Therefore, exponential decay of $P_{e_2}(\Lambda|\theta_2)$ is achieved for all $\theta_2 \in \Theta_2$, if and only if (39) holds for all $\theta_2 \in \Theta_2$.

It can be seen from (39) that the simple *necessary* condition for achieving exponential decay of the mis-detection probability for all $\theta_2 \in \Theta_2$ is given by

$$\lambda(\theta') \leq D(\mathbf{P}_2||P_{\theta_1'}) \triangleq \inf_{P_{\theta_2} \in \mathbf{P}_2} D(P_{\theta_2}||P_{\theta_1'}), \quad \forall(\theta') \in \Theta. \tag{40}$$

Note that this condition is also sufficient when Θ_2 is an open set, since for every θ_2, one can find $\delta(\theta_2) > 0$ such that $B(\theta_2, \delta(\theta_2)) \subseteq \Theta_2$.

In view of this result, for every θ_1, the value of $D(\mathbf{P}_2||P_{\theta_1})$ can be interpreted as the supremum over all error rates under H_1 that could be achieved by a universal decision rule, which still guarantees for all $\theta_2 \in \Theta_2$ exponential decay of the error probability under H_2. In the classical Neyman-Pearson approach, where two sources P_{θ_1} and P_{θ_2} are given, it is well-known (Stein's Lemma, see, e.g., [16, Corollary 1.2]) that the best exponential rate of the first kind error probability, when the second kind error probability is bounded away from 1, is $D(P_{\theta_2}||P_{\theta_1})$. Hence, our result can be seen as generalization of Stein's Lemma to composite hypothesis testing.

Theorem 2 can also be used for establishing conditions on the richness and the structure of the parameter sets under which universal efficient decision rules exist. Clearly, an efficient test exists if and only if there exists some *positive* $\lambda(\cdot)$ that satisfies (39) for all $\theta_2 \in \Theta_2$. For example, suppose that Θ_1 and Θ_2 are *separated away* in the sense that there exists some $\delta > 0$ such that

$$D(\mathbf{P}_2^\delta||\mathbf{P}_1) \triangleq \inf_{\theta_1 \in \Theta_1} D(\mathbf{P}_2^\delta||P_{\theta_1}) \triangleq \inf_{\theta_1 \in \Theta_1} \inf_{\mu \in \Theta_2^\delta} D(P_\mu||P_{\theta_1}) > 0, \tag{41}$$

where $\Theta_2^\delta \triangleq \bigcup_{\theta_2\in\Theta_2} B(\theta_2,\delta)$ is a δ-smoothing of Θ_2. Then, the threshold function $\lambda(\theta) = \xi D(\mathbf{P}_2^\delta||P_{\theta_1})$, with $0 < \xi \leq 1$, guarantees exponential decay of both error probabilities for all $\theta \in \Theta$. Moreover, in this case, for any constant threshold function $\lambda(\theta) \equiv \lambda_0$ with $\lambda_0 \leq D(\mathbf{P}_2^\delta||\mathbf{P}_1)$, equation (39) holds for all $\theta_2 \in \Theta_2$. Therefore, for every such λ_0, also the generalized Neyman-Pearson criterion leads to an efficient decision rule. However, this is a more conservative approach since for all $\theta_1 \in \Theta_1$ the first kind error exponent is only guaranteed to be λ_0, while in the competitive Neyman-Pearson approach, for those θ_1 that are "far" enough from Θ_2, we can guarantee much higher values of the first kind error exponent.

In certain problems of practical interest Θ_1 happens to be a subset of $\overline{\Theta}_2$ (e.g., $\Theta_1 = 1/2$ and $\Theta_2 = (0,1/2)\cup(1/2,1)$). For this geometry, $D(\mathbf{P}_2||P_{\theta_1})$ vanishes for all $\theta_1 \in \Theta_1$. As a result, an exponential decay of both error probabilities cannot be achieved simultaneously.

Another interesting situation is when Θ_1 and Θ_2 are open sets. As was mentioned earlier, if Θ_2 is open, then the necessary and sufficient condition for the exponential decay of the second kind error probability is $\lambda(\theta) \leq D(\mathbf{P}_2||P_{\theta_1})$. Hence, to prove the existence of an efficient test, it is enough to show that $D(\mathbf{P}_2||P_{\theta_1})$ is positive for all θ_1. Suppose, conversely, that $D(\mathbf{P}_2||P_{\theta_1}) = 0$ for some θ_1. Then, by continuity of $D(\cdot||\theta_1)$, there exists some $\mu \in \Theta_1$ that belongs to $\overline{\Theta}_2$. Since Θ_1 and Θ_2 are disjoint, this μ necessarily lies on the boundary of $\overline{\Theta}_2$ and hence any neighborhood of μ contains elements of Θ_2, contradicting the openness of Θ_1. Thus, for instance, for the threshold function $\lambda(\theta) = \xi D(\mathbf{P}_2||P_{\theta_1})$, with $0 < \xi \leq 1$, exponential decay of both error probabilities is guaranteed for all $\theta \in \Theta$. Note that even if $D(\mathbf{P}_2||\mathbf{P}_1) \triangleq \inf_{\theta_1\in\Theta_1} D(\mathbf{P}_2||P_{\theta_1})$ vanishes (and hence Θ_1 and Θ_2 are not separated away), e.g., if $\Theta_1 = (0,1/2)$ and $\theta_2 = (1/2,1)$, it is still possible to distinguish efficiently between the hypotheses, whereas the generalized Neyman-Pearson approach, in this situation, fails to universally achieve exponentially vanishing error probabilities, since no constant threshold $\lambda_0 > 0$ can satisfy (40).

Proof of Theorem 2. In this proof, we use large deviations techniques [21]. By definition of the set C

$$P_{e_2}(\Lambda^n|\theta_2) = p_{\theta_2}(y \in \Lambda_1^n) = p_{\theta_2}(Q_y \in C \cap \mathcal{L}_n), \tag{42}$$

where $\mathcal{L}_n$ denotes the set of all empirical measures induced by sequences of length n. Using Sanov's theorem for finite alphabets [21, Theorem 2.1.10], we have

$$\inf_{Q\in C} D(Q||P_{\theta_2}) \leq \liminf_{n\to\infty} -\frac{1}{n}\log p_{\theta_2}(Q_y \in C \cap \mathcal{L}_n) \leq \inf_{Q\in C^0} D(Q||P_{\theta_2}). \tag{43}$$

By definition of $g(\cdot)$ as the point-wise infimum of a family of continuous functions, it is upper semi-continuous, which implies that C is an open set. Therefore, an upper bound in (43) coincides with a lower bound and the part (a) is proved.

By continuity of $D(\cdot||P_{\theta_2})$, the error exponent under H_2, given by (36), vanishes if and only if $P_{\theta_2} \in \overline{C}$. Thus, $P_{e_2}(\Lambda^n|\theta_2)$ converges to zero exponentially fast iff $P_{\theta_2} \in \left(\overline{C}\right)^c = (C^c)^o$. This condition is equivalent to the following: There exists some $\delta(\theta_2)$ such that $B(\theta_2, \delta(\theta_2)) \subseteq C^c$, or equivalently, for every $\mu \in B(\theta_2, \delta(\theta_2))$,

$$g(P_\mu) = \inf_{\theta'\in\Theta}\left[D(P_\mu||P_{\theta_1'}) - \lambda(\theta')\right] \geq 0, \tag{44}$$

which completes the proof. $\square$

In general, to achieve exponential decay of both error probabilities, $\lambda(\theta)$ may be an arbitrary function of θ which satisfies (39) for all $\theta \in \Theta$. However, there are two specific choices mentioned earlier that would be advisable to examine more closely.

One, perhaps the most natural choice in view of Theorem 2, is $\lambda(\theta) = \xi D(\mathbf{P}_2||P_{\theta_1})$, where $0 < \xi < 1$. In this case, a decision rule is required to achieve only a certain fraction ξ of the maximal universally achievable exponent of the first kind error probability. It follows from (37) that when $\xi \to 0$, the set C tends to include only Q satisfying

$$\inf_{\theta_1' \in \Theta_1} D(Q||P_{\theta_1'}) = 0, \tag{45}$$

i.e., $Q \in \overline{\mathbf{P}}_1$. Therefore, by Theorem 2 the second kind error exponent (36) tends to be

$$e_2(\Lambda|\theta_2) = D(\mathbf{P}_1||P_{\theta_2}) \triangleq \inf_{Q \in \mathbf{P}_1} D(Q||P_{\theta_2}). \tag{46}$$

Similarly to $D(\mathbf{P}_2||P_{\theta_1})$, the value of $D(\mathbf{P}_1||P_{\theta_2})$ for every θ_2 can be interpreted as the supremum over all universally achievable error rates under H_2, for which the first kind error exponent does not vanish for all θ_1. Thus, the choice of $0 < \xi < 1$ controls the balance between false-alarm and mis-detection rates. At the upper edge of the range of ξ ($\xi \to 1$) the highest false-alarm rate $D(\mathbf{P}_2||P_{\theta_1})$ is attained, and at the lower edge ($\xi \to 0$), we obtain the highest mis-detection rate $D(\mathbf{P}_1||P_{\theta_2})$.

It should be stressed that, at least in the case that Θ_1 and Θ_2 are open sets, for any $0 < \xi < 1$, the exponential decay of both error probabilities is guaranteed for all sources in $\mathbf{P}_1$ and $\mathbf{P}_2$.

Another interesting choice of $\lambda(\theta)$ is $\lambda(\theta) = \xi E^*(\theta)$, where $E^*(\theta)$ is the error exponent function of the LRT associated with the Bayesian setting of the simple hypothesis testing problem and ξ is a given positive number (not necessary less than 1). The value of ξ, in this case, can be interpreted as the maximal tolerable level of the loss (or the minimal gain) in the first kind error rate relative to $E^*(\theta)$. By Theorem 2, the maximal ξ that guarantees exponential decay of the second kind error probability is upper bounded by $\inf_{\theta \in \Theta} \frac{D(\mathbf{P}_2||\theta_1)}{E^*(\theta)}$. As was mentioned in Introduction, this choice of the threshold function emphasizes the relation of our competitive Neyman-Pearson approach to the competitive minimax approach of [18]. Moreover, if we select ξ to be the largest possible number $\xi = \xi^*$ such that there exists a decision rule for which both error exponents are greater or equal to $\xi E^*(\theta)$, then the behavior of the error probabilities would be essentially symmetrical and the resulting test will be nearly optimal in the competitive minimax sense (with ξ^*) for the Bayesian setting.

More precisely, let

$$\Lambda_1^n(\gamma_n) \triangleq \left\{ \boldsymbol{y} : \inf_{\theta\in\Theta} \Big(D(Q_{\boldsymbol{y}}||P_{\theta_1}) - \xi^* E^*(\theta) \Big) < -\gamma_n \right\}, \tag{47}$$

where

$$\xi^* \triangleq \sup \left\{ \xi : \quad \liminf_{n\to\infty} -\frac{1}{n} \log \inf_{\Omega^n} \sup_{\theta\in\Theta} \frac{P_e(\Omega^n|\theta)}{2^{-n\xi E^*(\theta)}} \geq 0 \right\} \tag{48}$$

and $\gamma \triangleq \{\gamma_n\}_{n=1}^{\infty}$ is a positive sequence that decays to zero as $n \to \infty$ with sufficiently slow rate that will be specified later, in the proof of Corollary 3.

We assume that $\xi^* E^*(\theta)$ is a universally achievable error rate, that is, supremum in (48) can be replaced by maximum. If it is not, then the results to follow can be proved for any $(\xi^* - \epsilon)$, where $\epsilon > 0$ is an arbitrary small number.

Corollary 3. *Let the decision rule $\Lambda(\gamma)$ be defined as in (47), (48) . Then:*

$$\liminf_{n\to\infty} -\frac{1}{n} \log P_e(\Lambda^n(\gamma_n)|\theta) \geq \xi^* E^*(\theta), \quad \forall \theta \in \Theta. \tag{49}$$

The proof is based on the same technique as the proof of Theorem 1 and appears in Appendix I.

Note that although (47) and the test of [18] are different, the implementation of both these decision rules requires the exact value of ξ^*, which is usually hard to find. We next demonstrate another test, which does not use the knowledge of ξ^* and, nevertheless, uniformly achieves the maximal fraction ξ^* of the optimal error exponent.

Define

$$W_1^n(\gamma_n) \triangleq \left\{ \boldsymbol{y} : \inf_{\theta\in\Theta} \frac{D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \gamma_n}{E^*(\theta)} < \inf_{\theta\in\Theta} \frac{D(Q_{\boldsymbol{y}}||P_{\theta_2}) + \gamma_n}{E^*(\theta)} \right\}, \tag{50}$$

where γ_n is the same as in (47). The next theorem establishes the asymptotic optimality of $W(\gamma)$ in the competitive minimax sense.

Theorem 4. *Let the decision rule $W(\gamma)$ be defined as in (50). Then:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_e(W^n(\gamma_n)|\theta) \geq \xi^* E^*(\theta), \quad \forall\theta\in\Theta. \tag{51}$$

where ξ^* is defined by (48).

Proof of Theorem 4. Let us define $\widetilde{\Lambda}(\gamma)$ as

$$\widetilde{\Lambda}_2^n(\gamma_n) = \left\{ \boldsymbol{y} : \inf_{\theta\in\Theta} \Big(D(Q_{\boldsymbol{y}}||P_{\theta_2}) - \xi^* E^*(\theta) \Big) < -\gamma_n \right\}. \tag{52}$$

By definition of $W(\gamma)$, whatever the value of ξ^* is, the following condition is satisfied: For every $\boldsymbol{y} \in W_1^n(\gamma_n)$, either

$$\inf_{\theta\in\Theta} \frac{D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \gamma_n}{E^*(\theta)} < \xi^* \tag{53}$$

or

$$\inf_{\theta\in\Theta} \frac{D(Q_{\boldsymbol{y}}||P_{\theta_2}) + \gamma_n}{E^*(\theta)} \geq \xi^*. \tag{54}$$

Equivalently, either

$$\inf_{\theta\in\Theta} \left(D(Q_{\boldsymbol{y}}||P_{\theta_1}) - \xi^* E^*(\theta_1,\theta_2) \right) < -\gamma_n \tag{55}$$

or

$$\inf_{\theta\in\Theta} \left(D(Q_{\boldsymbol{y}}||P_{\theta_2}) - \xi^* E^*(\theta_1,\theta_2) \right) \geq -\gamma_n. \tag{56}$$

It means that $W_1^n(\gamma_n) \subseteq \Lambda_1^n(\gamma_n) \cup \widetilde{\Lambda}_1^n(\gamma_n)$. Employing the union bound, we obtain

$$P_{e_2}(W^n(\gamma_n)|\theta_2) = p_{\theta_2}(\boldsymbol{y} \in W_1^n(\gamma_n)) \leq p_{\theta_2}(\boldsymbol{y} \in \Lambda_1^n(\gamma_n) \cup \widetilde{\Lambda}_1^n(\gamma_n)) \tag{57}$$

$$\leq p_{\theta_2}(\boldsymbol{y} \in \Lambda_1^n(\gamma_n)) + p_{\theta_2}(\boldsymbol{y} \in \widetilde{\Lambda}_1^n(\gamma_n)). \tag{58}$$

Hence, the second kind error exponent associated with $W(\gamma)$ is lower-bounded by

$$e_2(W(\gamma)|\theta_2) \geq \min\left\{ e_2(\Lambda(\gamma)|\theta_2), e_2(\widetilde{\Lambda}(\gamma)|\theta_2) \right\} \geq \xi^* E^*(\theta), \quad \forall\theta\in\Theta, \tag{59}$$

where the last inequality follows from Corollary 3. Using similar arguments, it can also be shown that

$$e_1(W(\gamma)|\theta_1) \geq \xi^* E^*(\theta), \quad \forall \theta \in \Theta. \tag{60}$$

Thus, combining the last two equations, we have

$$\liminf_{n\to\infty} -\frac{1}{n} \log P_e(W^n(\gamma_n)|\theta) = \min\left\{e_1(W(\gamma)|\theta_1), e_2(W(\gamma)|\theta_2)\right\} \tag{61}$$

$$\geq \xi^* E^*(\theta), \quad \forall \theta \in \Theta, \tag{62}$$

and the theorem is proved.

□

Although merely i.i.d. sources were considered here, all our results can readily be extended to finite alphabet Markov sources. The only difference is in defining of the empirical entropy and the divergence, where the conditional empirical probabilities will be used instead of the ordinary ones.

It also extends to the class of FS AVSs with known state sequences. This could be useful in problems of communication across unknown channels, where a probability distribution of the channel output y_t at the time instant t depends on the input to the channel. In this case, the sequence of states associated with a hypothesis H_i is determined by the corresponding channel input sequence, which is known to the decoder.

In addition, similarly to [13], using a slightly weaker version of the optimality criterion, the generalization to the continuous alphabet case is possible. An optimal test, in this case, is based on the continuous version of the relative entropy and employs a "δ-smoothing" of decision regions (see [13]).

We remark that the method of types for i.i.d. sources [16], which, together with large deviations techniques [21], was used in this Section for the general analysis, can be extended to more general situations, such as Markov sources and unifilar FS sources in the discrete case [12], and exponential families [9] and, in particular, Gaussian models [22] in the continuous case. Hence, our approach can naturally be extended to all these important and commonly used parametric models.

Finally, our results can be generalized to the multiple hypothesis testing problem, where there are M composite hypotheses, provided that M does not grow exponentially with n. Unfortunately, using the same analysis technique, we were unable to extend them to the general case, where M grows exponentially with n, which in turn has a very important application in universal decoding for unknown communication channels.

3 Applications

This section is devoted to applications of our results and their extensions to some of the frequently encountered problems in communications, signal processing and detection theory areas. We investigate the usefulness of our approach in the context of these specific examples and compare it to the generalized Neyman-Pearson approach and other widely used methods.

One example is the problem of classifying an observation sequence into one of two unknown sources, when each source is represented by an independent training sequence. The optimal test in the generalized Neyman-Pearson sense, derived in [4], uses only one training sequence but is inconsistent. In contrast, we demonstrate that our approach leads to a universal test that employs both training sequences and guarantees exponential decay of the error probabilities for all distinct sources.

Another example is estimating the order of a finite-alphabet Markov source. It will be shown that, in this problem, the best universally achievable exponent of the overestimation probability vanishes, which implies that an efficient universal Markov order estimator does not exist at all. This fact was also proved in [19].

Finally, application of our approach to the problem of detection of signals transmitted across an unknown finite-alphabet finite-state channel is examined and compared to the generalized Neyman-Pearson approach of [11]. In addition, an optimal detector is explicitly derived for the class of Gaussian intersymbol interference (ISI) channels with finite input alphabet.

3.1 Classification with Training Sequences

The problem of classifying probabilistic information sources, whose statistics are only partially available though training sequences, is frequently encountered in speech recognition applications, signal detection and digital communications. This problem can be treated as one of the multiple composite hypothesis testing as follows.

Let $\{P_\phi : \phi \in \Phi\}$ be a certain parametric family of PMF's over a finite alphabet A. There are M distinct unknown sources, P_{ϕ_1}, P_{ϕ_2},..., P_{ϕ_M}, whose probability measures belong to this family, i.e. $\phi_i \in \Phi$, $i = 1, 2, \ldots, M$. We are given a test sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in A^n$ that must be classified as having been produced by one of the M sources. In addition, for each ϕ_i ($i = 1, 2, \ldots, M$), there is a training sequence $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im}) \in A^m$ emitted from the source P_{ϕ_i}. It is assumed that the training sequences are independent of each other and of the test sequence $\boldsymbol{x}$. The problem is to decide among hypotheses $\{H_i, i = 1, 2, \ldots, M\}$, where hypothesis H_i is that $\boldsymbol{x}$ and $\boldsymbol{t}_i$ originated from the same source. A decision rule Ω for this problem is a sequence of partitions of the observation space $A^n \times (A^m)^M$ into M disjoint regions $\Omega_1^n, \Omega_2^n, \ldots \Omega_M^n$, where the test sequence $\boldsymbol{x}$ is classified as coming from the source P_{ϕ_i} iff $(\boldsymbol{x}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_M) \in \Omega_i^n$.

As an important example, we consider the binary classification problem of memoryless sources. That is, we assume that $M = 2$ and $\{P_\phi : \phi \in \Phi\}$ is the class of all memoryless probability measures over the finite alphabet A with *strictly positive* letter probabilities. In addition, we assume that the asymptotic regime of the problem is such that the length of the training sequences m grows linearly with the length of the test sequence n, namely the ratio $r \triangleq \frac{m}{n}$ is constant for all large n. In this configuration, the conditional probability distribution of the entire data set $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{t}_1, \boldsymbol{t}_2)$ under H_i is given by

$$H_1 : \quad p_{\theta_1}(\boldsymbol{y}) = p_{\phi_1}(\boldsymbol{x}) p_{\phi_1}(\boldsymbol{t}_1) p_{\phi_2}(\boldsymbol{t}_2), \tag{63}$$

$$H_2 : \quad p_{\theta_2}(\boldsymbol{y}) = p_{\phi_2}(\boldsymbol{x}) p_{\phi_1}(\boldsymbol{t}_1) p_{\phi_2}(\boldsymbol{t}_2), \tag{64}$$

where $\theta_1 \triangleq (\phi_1, \phi_1, \phi_2) \in \Theta_1$ and $\theta_2 \triangleq (\phi_2, \phi_1, \phi_2) \in \Theta_2$ are unknown parameters. Note, first, that the parameter sets $\Theta_1 \triangleq \big\{(\phi_1, \phi_1, \phi_2) \in \Phi^3 : \phi_1 \neq \phi_2\big\}$ and $\Theta_2 \triangleq \big\{(\phi_2, \phi_1, \phi_2) \in \Phi^3 : \phi_1 \neq \phi_2\big\}$ are not separated away. Secondly, they are relatively open rather than open, i.e., Θ_i is open on the hyperplane $\big\{(\phi_i, \phi_1, \phi_2) \in \Phi^3\big\}$, $i = 1, 2$, but it is not open on Φ^3. In addition, θ_1 and θ_2 are related in the sense that θ_2 is completely specified by θ_1. This fact reduces the degree of uncertainty in the parameters and may potentially improve the performance of a universal test.

Let $\boldsymbol{Q}_{\boldsymbol{y}} \triangleq (Q_{\boldsymbol{x}}, Q_{\boldsymbol{t}_1}, Q_{\boldsymbol{t}_2})$ denote the triplet of empirical PMF's associated with $\boldsymbol{y}$. In this case, $\boldsymbol{Q}_{\boldsymbol{y}}$ is a sufficient statistic for asymptotically optimal classification in the error exponent sense. For any two triplets of PMF's $\boldsymbol{Q}_1 \triangleq (Q_{11}, Q_{12}, Q_{13})$ and $\boldsymbol{Q}_2 \triangleq (Q_{21}, Q_{22}, Q_{23})$, let us define the following functional:

$$\widetilde{D}(\boldsymbol{Q}_1 || \boldsymbol{Q}_2) \triangleq D(Q_{11} || Q_{21}) + rD(Q_{12} || Q_{22}) + rD(Q_{13} || Q_{23}). \tag{65}$$

Now, if we replace $D(\cdot||\cdot)$ by $\widetilde{D}(\cdot||\cdot)$ in the formulations of Theorems 1, 2 and 4, their results will hold for the above-defined classification problem. Thus, an optimal classifier in the competitive Neyman-Pearson sense can be written in the following form:

$$\begin{aligned} \Lambda_1^n = \Bigg\{ \boldsymbol{y} : \inf_{(\phi_1, \phi_2) \in \Phi^2} \Big(& D(Q_{\boldsymbol{x}} || P_{\phi_1}) + rD(Q_{\boldsymbol{t}_1} || P_{\phi_1}) \\ & + rD(Q_{\boldsymbol{t}_2} || P_{\phi_2}) - \lambda(\phi_1, \phi_2) \Big) < 0 \Bigg\}. \end{aligned} \tag{66}$$

Note that, in the particular case of $\lambda(\phi_1, \phi_2) = \lambda_0$, where λ_0 is a positive constant, this test coincides with the GLRT studied in [4],[7]. It was shown there that only one training sequence is needed to achieve an asymptotic optimality in the generalized Neyman-Pearson sense. To see it from (66), observe that $\lambda(\phi_1, \phi_2) = \lambda_0$ does not take part in the minimization and therefore the minimization over ϕ_2 sets to zero $D(Q_{\mathbf{t}_2}||P_{\phi_2})$. This emphasizes the pessimistic nature of the generalized Neyman-Pearson criterion, since the additional information associated with the second training sequence is ignored. As a result, no λ_0 can guarantee exponential decay of both error probabilities for all sources. In contrast, the decision rule (66) generally uses both training sequences and, as will be shown later, there exists $\lambda(\phi_1, \phi_2)$ that leads to an efficient classifier.

It follows from Theorem 2 that the upper bound on the efficient threshold function is given by

$$\widetilde{D}(\mathbf{P}_2||P_{\theta_1}) \triangleq \inf_{\theta'_2 \in \Theta_2} \widetilde{D}(P_{\theta'_2}||P_{\theta_1}) \tag{67}$$

$$= \inf_{(\phi'_1, \phi'_2) \in \Phi^2} D(P_{\phi'_2}||P_{\phi_1}) + rD(P_{\phi'_1}||P_{\phi_1}) + rD(P_{\phi'_2}||P_{\phi_2}) \tag{68}$$

$$= \inf_{\phi'_2 \in \Phi} D(P_{\phi'_2}||P_{\phi_1}) + rD(P_{\phi'_2}||P_{\phi_2}). \tag{69}$$

By a standard minimization technique, it can be shown that the minimum in the last expression is attained by $P_{\phi'_2} = P_{\phi_1 \times \phi_2}$, where $P_{\phi_1 \times \phi_2}$ denotes a normalized exponential combination of P_{ϕ_1} and P_{ϕ_2} defined by

$$P_{\phi_1 \times \phi_2}(\alpha) \triangleq \frac{P^s_{\phi_1}(\alpha) P^{1-s}_{\phi_2}(\alpha)}{\sum_{\alpha' \in A} P^s_{\phi_1}(\alpha') P^{1-s}_{\phi_2}(\alpha')}, \qquad \forall \alpha \in A, \tag{70}$$

where $s = \frac{1}{1+r}$. Hence, in this specific problem of classification with training sequences, the expression $D_c(P_{\phi_2}||P_{\phi_1}) \triangleq D(P_{\phi_1 \times \phi_2}||P_{\phi_1}) + rD(P_{\phi_1 \times \phi_2}||P_{\phi_2})$ measures the universal distinguishability between the composite hypotheses associated with the partially known sources P_{ϕ_1} and P_{ϕ_2}. It is an analogue to $D(P_{\phi_2}||P_{\phi_1})$, which has the similar role in the simple hypothesis testing (Stein's Lemma [16, Corollary 1.2]).

As can easily be verified, $D_c(P_{\phi_2}||P_{\phi_1})$ has the following properties:

$$1. \quad 0 \leq D_c(P_{\phi_2}||P_{\phi_1}) \leq D(P_{\phi_2}||P_{\phi_1}), \quad \forall r \geq 0, \quad \forall(\phi_1,\phi_2) \in \Phi^2, \tag{71}$$

$$2. \quad \lim_{r\to\infty} D_c(P_{\phi_2}||P_{\phi_1}) = D(P_{\phi_2}||P_{\phi_1}), \quad \forall(\phi_1,\phi_2) \in \Phi^2, \tag{72}$$

where equality holds on the left-hand side of (71) if and only if $\phi_1 = \phi_2$ or $r = 0$. Now, it can be observed that when r goes to infinity, i.e., the training sequences are considerably longer than the test sequence, our universal test Λ, which is independent of P_{ϕ_1} and P_{ϕ_2}, performs as well as the optimal LRT for that P_{ϕ_1} and P_{ϕ_2} in the sense of Stein's Lemma. On the other hand, if r tends to zero, then the best universally achievable false-alarm rate vanishes. This result is similar to that presented in [7], where it was shown that no universal classifier can perform efficiently unless the training sequences length increases at least linearly with the classified sequence length. Finally, for any $0 < r < \infty$, $D_c(P_{\phi_2}||P_{\phi_1})$ is positive for all $\phi_1 \neq \phi_2$. Hence, following the discussion after Theorem 2, it can be shown that for the threshold function $\lambda(\phi_1,\phi_2) = \xi D_c(P_{\phi_2}||P_{\phi_1})$, with $0 < \xi < 1$, exponential decay of both error probabilities will be guaranteed for all distinct sources.

For the M-hypothesis problem, our decision rule can be extended to a rejection decision scheme. This scheme is allowed not to make a decision (and request another test sequence for an additional attempt). In this case, the union of all decision regions is not equal to the entire observation space and $\Omega_R^n \triangleq \left(\bigcup_{i=1}^M \Omega_i^n\right)^c$ is called the rejection zone. If $(\boldsymbol{x},\boldsymbol{t}_1,\ldots,\boldsymbol{t}_M) \in \Omega_R^n$, then a rejection is made. The probability of error under H_i is given by

$$P_{e_i}(\Omega^n|\phi) = \sum_{\boldsymbol{y}\in\bigcup_{j\neq i}\Omega_j^n} p_{\phi_i}(\boldsymbol{x}) \prod_{k=1}^M p_{\phi_k}(\boldsymbol{t}_k), \tag{73}$$

where $\phi \triangleq (\phi_1,\ldots,\phi_M) \in \Phi^M$ and $\boldsymbol{y} \triangleq (\boldsymbol{x},\boldsymbol{t}_1,\ldots,\boldsymbol{t}_M)$.

We are interested in a decision rule, which is optimal in the sense that it minimizes an exponential rate of the rejection probability subject to the constraint that all error probabilities decay exponentially in n with rate at least $\lambda(\phi)$.

Consider the test statistic

$$f_i(\boldsymbol{y}) = \inf_{\phi \in \Phi^M} \left(D(Q_{\boldsymbol{x}}||P_{\phi_i}) + r \sum_{k=1}^{M} D(Q_{\boldsymbol{t}_k}||P_{\phi_k}) - \lambda(\phi) \right), \quad i = 1, \ldots, M, \tag{74}$$

and let the classifier Λ be defined by

$$\begin{aligned} \Lambda_1^n &= \left\{ \boldsymbol{y} : f_k(\boldsymbol{y}) \geq 0, \quad \forall k = 2, \ldots, M \right\}, \\ \Lambda_i^n &= \left\{ \boldsymbol{y} : f_i(\boldsymbol{y}) < 0, f_k(\boldsymbol{y}) \geq 0, \quad \forall k \neq i, 1 \leq k \leq M \right\}, \quad \forall i = 2, \ldots, M. \end{aligned} \tag{75}$$

Similarly to the proof of [4, Theorem 2], it can be demonstrated that Λ is optimal under the competitive Neyman-Pearson criterion with rejection. Note that this decision rule reduces to the GLRT, that was developed in [4] for the rejection scheme, if $\lambda(\phi)$ is a constant threshold. For this setting, the GLRT uses all training sequences, but still it does not assure an exponential decay of the rejection probability. In contrast, combining the results of Theorem 2 and [4, Theorem 3], it can be shown that the universal classifier (75) is efficient, provided that

$$\lambda(\phi) < \min_{\substack{k,j \in \{1,\ldots,M\} \\ k \neq j}} D_c(P_{\phi_j}||P_{\phi_k}), \tag{76}$$

for all $\phi \in \Phi^M$.

To gain some intuition regarding the way the information carried by training sequences can be efficiently used by a universal test to control the rejection rate, consider the following decision procedure: The parameters of the sources are first estimated from the given training data. Then, based on these estimates, a suitable threshold λ is chosen. Finally, this threshold and the estimates of the parameters are used in the LRT for the simple hypothesis testing problem with rejection. This approach, with an appropriate strategy of choosing λ, can be shown to universally achieve an exponential decay of the rejection probability, but it might not be optimal in the error exponent sense. It is interesting to point out, however, that in this decision scheme, a good value of λ depends, via the training sequences, on the unknown parameters, which is similar to the idea of variable threshold in the competitive Neyman-Pearson approach.

For the Bayesian setting, using the optimality of Λ in the competitive Neyman-Pearson sense with rejection, one can readily extend Theorem 4 to the case of M hypotheses. Thus, an asymptotically optimal test in the competitive minimax sense (with ξ^*) will classify the test sequence $\boldsymbol{x}$ as being generated by the source P_{ϕ_i} for which

$$g_i(\boldsymbol{y}) \triangleq \inf_{\phi\in\Phi^M} \frac{D(Q_{\boldsymbol{x}}||P_{\phi_i}) + r\sum_{k=1}^{M} D(Q_{\boldsymbol{t}_k}||P_{\phi_k}) + \gamma_n}{E^*(\phi)}. \tag{77}$$

is minimal. It should be pointed out that if $r \to \infty$, then P_{ϕ_k} that minimizes the r.h.s. of (77) is very close to $Q_{\boldsymbol{t}_k}$, $k = 1, \ldots, M$. Therefore, the decision rule (77) can be approximated by

$$\widehat{g}_i(\boldsymbol{y}) = D(Q_{\boldsymbol{x}}||Q_{\boldsymbol{t}_i}). \tag{78}$$

In other words, if the amount of training data is relatively large, then (77) performs nearly as "plug-in" (PI) method, where the ML estimates of the parameters, calculated from relative frequencies of letters in the training sequences, are plugged into the LRT.

3.2 Model Order Estimation

In this subsection, we apply our method to the estimation of the order k of a discrete-time finite-alphabet ergodic Markov source, when an upper bound k_0 on the true order is available. In [8], where this problem was studied in detail, an asymptotically optimal order estimator was developed in the sense of minimizing the underestimation probability while keeping the overestimation probability exponent at a certain level λ. This is a generalized version of the Neyman-Pearson criterion. It was shown in [8], that in contrast to other earlier proposed estimation algorithms, which achieve only subexponential decay of the overestimation probability, for every Markov source there exists $\lambda > 0$ such that both overestimation and underestimation probabilities vanish exponentially fast. The major deficiency of this approach, however, lies in the fact, as was explained earlier, that once $\lambda > 0$ was fixed, one can find a Markov source for which the underestimation probability tends to 1 [8, Remark 1]. Thus, an appealing issue is to examine whether our competitive Neyman-Pearson approach can remedy this problem.

To facilitate the discussion we assume that the true order is bounded by $k_0 = 1$. That is, it is desired to test the hypothesis H_1: an observed sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ was emitted from an unknown i.i.d. source, against the alternative H_2: $\boldsymbol{y}$ was emitted from an unknown first order Markov source. We also assume here that all *transition probabilities* of Markov sources are *strictly positive*. Thus, in this example, $\mathbf{P}_1 = \{P_{\theta_1}, \theta_1 \in \Theta_1\}$ is the class of all memoryless sources over a finite alphabet A with strictly positive letter probabilities, and $\mathbf{P}_2 = \{P_{\theta_2}, \theta_2 \in \Theta_2\}$ is the class of all stationary ergodic first order Markov sources over the alphabet A with strictly positive transition probabilities, which can not be reduced to memoryless sources. Observe that, in this setup, Θ_2 is open and $\Theta_1 \subseteq \overline{\Theta}_2$.

Let $s_i \triangleq y_{i-1} \in A$ denote the state of the Markov source at time instant i ($s_1 = y_0$ is the fixed initial state). Define the empirical joint probability of the letter $\alpha \in A$ and the state $s \in A$ in the vector $\boldsymbol{y} \in A^n$ as

$$q^1_{\boldsymbol{y}}(\alpha, s) \triangleq \frac{1}{n} \sum_{j=1}^{n} \delta(y_j = \alpha, s_j = s), \tag{79}$$

where $\delta(y_j = \alpha, s_j = s)$ is the indicator function for $y_j = \alpha$ jointly with $s_j = s$. The matrix $Q^1_{\boldsymbol{y}} \triangleq \left\{q^1_{\boldsymbol{y}}(\alpha, s) : \alpha \in A, s \in A\right\}$ can be viewed as an empirical first order Markov distribution associated with $\boldsymbol{y}$. Also, let

$$q^1_{\boldsymbol{y}}(s) = \sum_{\alpha \in A} q^1_{\boldsymbol{y}}(\alpha, s), \tag{80}$$

$$q^1_{\boldsymbol{y}}(\alpha|s) = \frac{q^1_{\boldsymbol{y}}(\alpha, s)}{q^1_{\boldsymbol{y}}(s)}, \tag{81}$$

where $q^1_{\boldsymbol{y}}(\alpha|s)$ is set to zero if the denominator $q^1_{\boldsymbol{y}}(s)$ vanishes. Next, define the divergence between $Q^1_{\boldsymbol{y}}$ and $P_{\theta_1} \in \mathbf{P}_1$ as

$$D(Q^1_{\boldsymbol{y}}||P_{\theta_1}) \triangleq \sum_{s \in A} q^1_{\boldsymbol{y}}(s) \sum_{\alpha \in A} q^1_{\boldsymbol{y}}(\alpha|s) \log \frac{q^1_{\boldsymbol{y}}(\alpha|s)}{p_{\theta_1}(\alpha)}. \tag{82}$$

Similarly, we define the divergence between the first order Markov source $P_{\theta_2} \in \mathbf{P}_2$ and the memoryless source $P_{\theta_1} \in \mathbf{P}_1$:

$$D(P_{\theta_2}||P_{\theta_1}) = \sum_{s \in A} p_{\theta_2}(s) \sum_{\alpha \in A} p_{\theta_2}(\alpha|s) \log \frac{p_{\theta_2}(\alpha|s)}{p_{\theta_1}(\alpha)}, \tag{83}$$

where $\{p_{\theta_2}(s), s \in A\}$ is an invariant probability measure associated with P_{θ_2}, i.e., $\sum_{s' \in A} p_{\theta_2}(s') p_{\theta_2}(s|s') = p_{\theta_2}(s)$.

Now, extending Theorem 1 to this case, we straightforwardly obtain that an optimal estimator in the competitive Neyman-Pearson sense will select an order $\widehat{k} = 0$ iff

$$\boldsymbol{y} \in \Lambda^n_1 = \left\{ \boldsymbol{y} : \inf_{\theta \in \Theta} \big(D(Q^1_{\boldsymbol{y}}||P_{\theta_1}) - \lambda(\theta) \big) < 0 \right\}. \tag{84}$$

Since Θ_2 is open, the necessary and sufficient condition on the threshold function $\lambda(\theta)$ under which an exponential decay of the underestimation probability is achieved for all first order Markov sources is given by (40):

$$\lambda(\theta) \leq D(\mathbf{P}_2||P_{\theta_1}), \quad \forall \theta \in \Theta. \tag{85}$$

Obviously, in this case, $D(\mathbf{P}_2||P_{\theta_1}) \equiv 0$. It means that no threshold function can assure that both overestimation and underestimation probabilities will vanish exponentially fast for all possible sources. Moreover, since our approach is optimal in the error exponents sense, it implies that an efficient universal estimation of the order of a Markov chain is not feasible at all. Naturally, if we relax the assumption associated with positivity of the transition probabilities, we would only decrease prior knowledge about the true underling model and definitely would not be able to construct an efficient estimator.

The reason for the nonexistence of an efficient universal test, in this particular case, is that the parameter set Θ_1 is a subset of the closure of Θ_2. In general, we may conclude that if for a certain $P_{\theta_1} \in \mathbf{P}_1$ one can find $P_{\theta_2} \in \mathbf{P}_2$ which is arbitrary close to P_{θ_1} in the informational divergence sense, then there is no possibility to efficiently distinguish between the hypotheses. Several other well-known examples fall in this category of the composite hypothesis testing problems: (i) discrimination problem, where we wish to decide whether or not two given sequences of random variables were emitted from the same source [4], [7], (ii) testing for independence - whether two sequences are mutually independent [5], [6], and (iii) testing for randomness - whether or not a given sequence consists of i.i.d. random variables [5], [6]. In all these examples the parameter sets Θ_1 and Θ_2 have the property described above and hence efficient universal tests do not exist.

3.3 Detection of Messages via Unknown Channels

Detection of signals transmitted across an unknown noisy communication channel is the very important problem in composite hypothesis testing. Some typical examples are the following: (i) radar target detection, where an unknown attenuation and phase shift are introduced by the channel in the transmitted signal [2], [23], (ii) identification problem and watermark detection (see e.g., [24], [25] and references therein), and (iii) digital communication over an unknown channel [16], [22], [26], [27].

In our framework, this problem is defined as follows. Consider an unknown channel W_ψ from some family of channels $\mathcal{W}$ defined by the conditional PMF's $\{w_\psi(\boldsymbol{y}|\boldsymbol{x}), \psi \in \Psi\}$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ is the channel input, $\boldsymbol{y} =$

$(y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n$ is the channel output, ψ is the index of the channel in the family, and Ψ is some index set. A transmitter uses a set $\mathcal{C} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M\}$ of M messages $\boldsymbol{x}^i \in \mathcal{X}^n$, $i = 1, 2, \ldots, M$, where M is a *fixed* positive integer, to send information across the channel. Given a received sequence $\boldsymbol{y}$ and the signal set $\mathcal{C}$, the decoder has to decide which of M possible messages was transmitted. In radar, identification and certain watermarking applications, however, the receiver is not required to carry out full decoding and needs only to decide whether or not an output sequence $\boldsymbol{y}$ corresponds to a particular input sequence. In this case, the problem essentially reduces to binary detection, where it is natural to use the Neyman-Pearson criterion to balance appropriately the trade-off between false-alarm and mis-detection rates.

The conditional distribution of $\boldsymbol{y}$ under hypothesis H_i ($\boldsymbol{x}^i$ was transmitted) is given by $p_{\theta_i}(\boldsymbol{y}) = w_\psi(\boldsymbol{y}|\boldsymbol{x}^{\boldsymbol{i}})$, where θ_i designates an unknown parameter associated with the hypothesis H_i, $i = 1, 2, \ldots, M$. Note that, as in the classification problem (Subsection 3.1), θ_i ($i = 1, 2, \ldots, M$) are related each to other. This relation, which affects the performance of a universal detector, is through the common channel parameter ψ and the structure of the signal set $\mathcal{C}$. Therefore, an additional interesting issue that arises here, is the choice of the signals that are suitable for universal detection. In the sequel, we provide an interesting characteristic of a good universally detectable signal set.

The detection problem defined above was studied in [11] for the class of finite-state channels over finite input and output alphabets. A universal decision rule, which is based on Lempel-Ziv (LZ) algorithm for source coding [28], was derived and shown to be asymptotically optimal in the generalized Neyman-Pearson sense. However, no results concerning the behavior of the mis-detection probability have been presented.

To describe the universal detector that is derived from our competitive Neyman-Pearson approach, let us assume, for the sake of simplicity, that $M = 2$ and $\mathcal{W}$ is the class of all finite-alphabet discrete memoryless channels (DMC's) with the transition probability function of the form

$$w_\psi(\boldsymbol{y}|\boldsymbol{x}) = \prod_{j=1}^{n} w_\psi(y_j|x_j). \tag{86}$$

In order to extend Theorem 2 to this case, we need the assumption that all transition probabilities are positive, namely, $w_\psi(y|x) > 0$ for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$ and $W_\psi \in \mathcal{W}$.

Let $z_i \triangleq (x_i^1, x_i^2) \in \mathcal{Z} \triangleq \mathcal{X}^2$, $i = 1, \ldots, n$, and $\boldsymbol{z} = (z_1, \ldots, z_n)$. Let $Q_{\boldsymbol{y},\boldsymbol{z}} \triangleq \{q_{\boldsymbol{y},\boldsymbol{z}}(y,z) : y \in \mathcal{Y}, z \in \mathcal{Z}\}$ denote the empirical joint PMF associated with the vectors $\boldsymbol{y}$ and $\boldsymbol{z}$, where

$$q_{\boldsymbol{y},\boldsymbol{z}}(y,z) \triangleq \frac{1}{n}\sum_{j=1}^{n} \delta(y_j = y, z_j = z), \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{87}$$

It can be shown that, in this case, $Q_{\boldsymbol{y},\boldsymbol{z}}$ serves as sufficient statistics for asymptotically optimal detection. Similarly, $Q_{\boldsymbol{y},\boldsymbol{x}^i}$ will denote the empirical joint PMF associated with $(\boldsymbol{y}, \boldsymbol{x}^i)$, $i = 1, 2$. Let $Q_{\boldsymbol{x}^i}$, $Q_{\boldsymbol{y}}$ and $Q_{\boldsymbol{z}}$ denote the empirical PMF associated with $\boldsymbol{x}^i$, $\boldsymbol{y}$ and $\boldsymbol{z}$, respectively. It will be assumed that $Q_{\boldsymbol{z}}$ is independent of the input sequences length n and will be referred to as the empirical joint probability distribution of the signal set. Also, define for all $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$

$$q_{\boldsymbol{y},\boldsymbol{z}}(y|z) \triangleq \begin{cases} q_{\boldsymbol{y},\boldsymbol{z}}(y,z)/q_{\boldsymbol{z}}(z), & q_{\boldsymbol{z}}(z) > 0 \\ 0, & q_{\boldsymbol{z}}(z) = 0 \end{cases} \tag{88}$$

Similarly, $q_{\boldsymbol{y},\boldsymbol{x}^i}(y|\alpha_i)$ will denote the empirical conditional probability of $y \in \mathcal{Y}$ given $\alpha_i \in \mathcal{X}$ corresponding to the empirical joint probability distribution of $(\boldsymbol{y}, \boldsymbol{x}^i)$. Finally, let us define the conditional divergence

$$D(Q_{\boldsymbol{y},\boldsymbol{z}}||W_\psi|Q_{\boldsymbol{z}}, \boldsymbol{x}^i) \triangleq \sum_{z=(\alpha_1,\alpha_2)\in\mathcal{Z}} q_{\boldsymbol{z}}(z) \sum_{y\in\mathcal{Y}} q_{\boldsymbol{y},\boldsymbol{z}}(y|z) \log \frac{q_{\boldsymbol{y},\boldsymbol{z}}(y|z)}{w_\psi(y|\alpha_i)}, \quad i = 1, 2. \tag{89}$$

An optimal detector in this case turns out to be the following: Decide that $\boldsymbol{x}^1$ has been transmitted iff

$$\boldsymbol{y} \in \Lambda_1^n \triangleq \left\{ \inf_{\psi\in\Psi} \left(D(Q_{\boldsymbol{y},\boldsymbol{z}}||W_\psi|Q_{\boldsymbol{z}}, \boldsymbol{x}^1) - \lambda(\psi) \right) < 0 \right\}, \tag{90}$$

where $\lambda(\psi)$ designates, for every $\psi \in \Psi$, the prescribed value of the false-alarm rate. Observe that

$$D(Q_{\boldsymbol{y},\boldsymbol{z}}||W_\psi|Q_{\boldsymbol{z}},\boldsymbol{x}^1) = I(Q_{\boldsymbol{x}^2}, Q_{\boldsymbol{y},\boldsymbol{x}^2}|Q_{\boldsymbol{x}^1}) + D(Q_{\boldsymbol{y},\boldsymbol{x}^1}||W_\psi|Q_{\boldsymbol{x}^1}), \tag{91}$$

where $I(Q_{\boldsymbol{x}^2}, Q_{\boldsymbol{y},\boldsymbol{x}^2}|Q_{\boldsymbol{x}^1})$ is the empirical conditional mutual information given by

$$I(Q_{\boldsymbol{x}^2}, Q_{\boldsymbol{y},\boldsymbol{x}^2}|Q_{\boldsymbol{x}^1}) \triangleq \sum_{z=(\alpha_1,\alpha_2)\in\mathcal{Z}} q_{\boldsymbol{z}}(z) \sum_{y\in\mathcal{Y}} q_{\boldsymbol{y},\boldsymbol{z}}(y|z) \log \frac{q_{\boldsymbol{y},\boldsymbol{z}}(y|z)}{q_{\boldsymbol{y},\boldsymbol{x}^1}(y|\alpha_1)}, \tag{92}$$

and

$$D(Q_{\boldsymbol{y},\boldsymbol{x}^1}||W_\psi|Q_{\boldsymbol{x}^1}), \triangleq \sum_{\alpha_1\in\mathcal{X}} q_{\boldsymbol{x}^1}(\alpha_1) \sum_{y\in\mathcal{Y}} q_{\boldsymbol{y},\boldsymbol{x}^1}(y|\alpha_1) \log \frac{q_{\boldsymbol{y},\boldsymbol{x}^1}(y|\alpha_1)}{w_\psi(y|\alpha_1)}. \tag{93}$$

It follows then, that in the case of $\lambda(\psi) = \lambda_0$, our decision rule (90) reduces to the one that compares $I(Q_{\boldsymbol{x}^2}, Q_{\boldsymbol{y},\boldsymbol{x}^2}|Q_{\boldsymbol{x}^1})$ to the threshold. As could be expected, it essentially coincides, in the special case of the memoryless channels, with the test proposed in [11].

Applying the result of Theorem 2 to the discussed problem, we obtain that the best universally achievable false-alarm rate that guarantees exponential decay of the mis-detection probability is upper bounded by

$$\lambda_{max}(\psi, Q_{\boldsymbol{z}}) \triangleq \inf_{\psi'\in\Psi} \sum_{z=(\alpha_1,\alpha_2)\in\mathcal{Z}} q_{\boldsymbol{z}}(z) \cdot D(W_{\psi'}(\cdot|\alpha_2)||W_\psi(\cdot|\alpha_1)). \tag{94}$$

As can be seen, $\lambda_{max}(\psi, Q_{\boldsymbol{z}})$ depends on the empirical joint probability distribution of the signal set $Q_{\boldsymbol{z}}$. In a sense, $\lambda_{max}(\psi, Q_{\boldsymbol{z}})$ measures the suitability of the signal set for universal detection. For example, suppose that all the components of the sequence $\boldsymbol{x}^1$ are equal to some $\alpha \in \mathcal{X}$ and all the components of the sequence $\boldsymbol{x}^2$ are equal to another letter $\beta \in \mathcal{X}$. Then, it can be observed from (94), that $\lambda_{max}(\psi, Q_{\boldsymbol{z}})$ vanishes for all $\psi \in \Psi$. It means that this set of signals, which may be good for a single known channel, is totally useless for universal detection. It would be reasonable, therefore, to examine the signal set $Q^*_{\boldsymbol{z}}(\psi)$ that attains $\max_{Q_{\boldsymbol{z}}} \lambda_{max}(\psi, Q_{\boldsymbol{z}})$. If $Q^*_{\boldsymbol{z}}(\psi)$ happens to be independent of ψ, this signal set would universally achieve the highest false-alarm rate, uniformly over the class of channels, while the mis-detection probability decays exponentially to zero. If, however, $Q^*_{\boldsymbol{z}}(\psi)$ depends on ψ, then another features of good signal sets have to be sought. As an example, the class of binary symmetrical channels (BSC's) is analyzed in Appendix II. It is demonstrated there that orthogonal signals are optimal in the above sense, i.e., the orthogonal signal set maximizes $\lambda_{max}(\psi, Q_{\boldsymbol{z}})$ uniformly over all BSC's.

In the applications that require full decoding at the receiver end, it is more appropriate to use the Bayesian setting of the detection problem. In this case, our asymptotically optimal decoder in the competitive minimax sense (with ξ^*) will select the massage $\boldsymbol{x}^i$ that minimizes

$$\inf_{\psi \in \Psi} \frac{D(Q_{\boldsymbol{y},\boldsymbol{z}} || W_\psi | Q_{\boldsymbol{z}}, \boldsymbol{x}^i) + \gamma_n}{E^*(\psi)}, \tag{95}$$

where $E^*(\psi)$ is the error exponent of the optimal ML decoder.

Although only DMC's where considered here, all our results straightforwardly extend to the case in which $\mathcal{W}$ is a family of finite-alphabet, finite-state channels with deterministic transitions and a fixed initial state. These channels are commonly used for modeling of the ISI channels. A finite-state channel $W_\psi \in \mathcal{W}$ is characterized by the following conditional probability distribution:

$$w_\psi(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} w_\psi(y_i|x_i, s_i), \tag{96}$$

where $s_i \in \mathcal{S}$ is a state of the channel at the time instant i and $w_\psi(y_i|x_i, s_i)$ is the probability of the current output of the channel $y_i \in \mathcal{Y}$, given the current input to the channel $x_i \in \mathcal{X}$ and the state $s_i \in \mathcal{S}$. The initial state s_1 is assumed fixed and known, and s_{i+1} is given by a deterministic next state function $g(x_i, s_i)$. In this case, the state sequence $\boldsymbol{s} = (s_1, \ldots, s_n)$ is determined by the channel input $\boldsymbol{x}$ and the initial state s_1. Again, we assume that $w_\psi(y|x, s) > 0$ for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$, $s \in \mathcal{S}$ and $W_\psi \in \mathcal{W}$.

For this class of channels, we yield the same detectors as in (90) and (95), but with $\boldsymbol{z} = (z_1, \ldots, z_n)$ being defined by

$$z_i \triangleq \left((x_i^1, s_i^1), (x_i^2, s_i^2)\right) \in \mathcal{Z} \triangleq (\mathcal{X} \times \mathcal{S})^2, \quad i = 1, \ldots, n, \tag{97}$$

where $\boldsymbol{s}^j = (s_1^j, \ldots, s_n^j)$ is the state sequence corresponding to the input sequence $\boldsymbol{x}^j$, $j = 1, 2$.

As mentioned at the end of Section 2, our general approach, with a slightly modified criterion of the optimality, can be extended to the infinite-alphabet case, where $D(\cdot||\cdot)$ is essentially replaced by its continuous version $I(\cdot||\cdot)$ (see e.g., [13]). In the context of the discussed detection problem, the corresponding generalization is possible to channels with a finite input alphabet and an infinite output alphabet. In this continuous case, the analogue to the assumption about positivity of the letter probabilities is that the support of the conditional probability density functions (PDF's) of the channel output given the channel input is the same for every input and for all channels in the family, where the support of the conditional PDF $w_\psi(\cdot|x)$ is the set of all $y \in \mathcal{Y}$ for which $w_\psi(y|x) > 0$. As an important and interesting example, we next consider the detection problem over the Gaussian ISI channel. Obviously, in the Gaussian case, the above assumption is trivially satisfied since the support of the Gaussian PDF is the whole real line $\mathbb{R}$.

Example. Consider the discrete-time Gaussian ISI channel characterized by

$$y_t = \sum_{j=0}^{\mathcal{L}} h_j x_{t-j} + n_t, \tag{98}$$

where $\{x_t\}$ is the input sequence from a finite alphabet $\mathcal{X}$, $\boldsymbol{h} = (h_0, \ldots, h_{\mathcal{L}})$ is the vector of unknown ISI coefficients, $\{n_t\}$ is zero-mean, Gaussian white noise with unit variance, and $\{y_t\}$ is the output sequence. The conditional PDF of the output $\boldsymbol{y}$ given the input $\boldsymbol{x}$ and parameterized by ISI sequence $\boldsymbol{h}$ is given by

$$w_{\boldsymbol{h}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \prod_{t=1}^{n} \exp\left\{-\frac{1}{2}\left(y_t - \sum_{j=0}^{\mathcal{L}} h_j x_{t-j}\right)^2\right\}, \tag{99}$$

where we assume that x_t is equal to an arbitrary but fixed letter $\alpha \in \mathcal{X}$ for $t \le 0$. Clearly, this channel is a finite-state channel with deterministic transitions and a fixed initial state. We again use the notation $z_t = \left((x_t^1, s_t^1), (x_t^2, s_t^2)\right) \in \mathcal{Z}$, where s_t^i is defined here as $s_t^i \triangleq (x_{t-\mathcal{L}}^i, \dots, x_{t-1}^i) \in \mathcal{S} = \mathcal{X}^{\mathcal{L}}$, $i = 1, 2$, and $\mathcal{Z} = (\mathcal{X} \times \mathcal{S})^2$.

Since the variance of the Gaussian noise is known, only the conditional empirical mean of $\boldsymbol{y}$ is essential for an asymptotically optimal decision. More precisely, let

$$\overline{y}_z \triangleq \frac{\sum_{t=1}^n y_t \cdot \delta(z_t = z)}{\sum_{t=1}^n \delta(z_t = z)}, \quad z \in \mathcal{Z} \tag{100}$$

i.e., $\overline{y}_z$ denotes the empirical mean of all y_t for which $z_t = z$. Then, the Gaussian empirical conditional PDF associated with $(\boldsymbol{y}, \boldsymbol{z})$

$$q_{\boldsymbol{y},\boldsymbol{z}}(y|z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - \overline{y}_z)^2\right\}, \quad y \in \mathbb{R}, z \in \mathcal{Z} \tag{101}$$

can be thought of as sufficient statistics.

It is well-known (and not difficult to show) that the divergence between two Gaussian distributions Q and P with means μ_q and μ_p, respectively, and unit variance is given by

$$I(Q||P) \triangleq \int_{-\infty}^{+\infty} q(y) \log \frac{q(y)}{p(y)} \, \mathrm{d}y = \frac{1}{2}(\mu_q - \mu_p)^2. \tag{102}$$

Therefore, the analogue to our asymptotically optimal detectors, in this continuous case, will be based on

$$I(Q_{\boldsymbol{y},\boldsymbol{z}}||W_{\boldsymbol{h}}|Q_{\boldsymbol{z}}, \boldsymbol{x}^i) \triangleq \sum_{z=(\boldsymbol{v}^1,\boldsymbol{v}^2)\in\mathcal{Z}} q_{\boldsymbol{z}}(z) \cdot \frac{1}{2}\left(\overline{y}_z - \sum_{j=0}^{\mathcal{L}} h_j v_j^i\right)^2, \quad i = 1, 2, \tag{103}$$

where $\boldsymbol{v}^i \triangleq (v_{\mathcal{L}}^i, \dots, v_0^i) \in \mathcal{X} \cdot \mathcal{S} = \mathcal{X}^{\mathcal{L}+1}$. Specifically, let $\boldsymbol{u} = (u_1, \dots, u_n) \triangleq \boldsymbol{x}^1 - \boldsymbol{x}^2$. The error exponent of the optimal ML decoder, which uses the knowledge of the ISI coefficients, is given by $\frac{1}{8}\sum_{t=1}^n \left(\sum_{j=0}^{\mathcal{L}} h_j u_{t-j}\right)^2$. Thus, the test statistic of our asymptotically optimal detector in the competitive minimax sense takes approximately the following form (ignoring γ_n):

$$\inf_{\boldsymbol{h}} \frac{\sum_{z=(\boldsymbol{v}^1,\boldsymbol{v}^2)\in\mathcal{Z}} q_{\boldsymbol{z}}(z)\left(\overline{y}_z - \sum_{j=0}^{\mathcal{L}} h_j v_j^i\right)^2}{\sum_{t=1}^{n}\left(\sum_{j=0}^{\mathcal{L}} h_j u_{t-j}\right)^2}. \tag{104}$$

Note that in the particular case of $\boldsymbol{h} = h_0$, i.e., $W_{\boldsymbol{h}}$ is the Gaussian memoryless channel with an unknown fading parameter h_0, the above test statistic simplifies to

$$\inf_{h_0} \frac{\sum_{z=(\alpha_1,\alpha_2)\in\mathcal{X}^2} q_{\boldsymbol{z}}(z)\left(\overline{y}_z - h_0\alpha_i\right)^2}{h_0^2\sum_{t=1}^{n} u_t^2}. \tag{105}$$

Since $\left(\sum_{t=1}^{n} u_t^2\right)$ is independent of h_0, defining $b \triangleq 1/h_0$ we yield that this detector selects $\boldsymbol{x}^i$ that minimizes

$$\inf_b \sum_{z=(\alpha_1,\alpha_2)\in\mathcal{X}^2} q_{\boldsymbol{z}}(z)\left(\alpha_i - b\overline{y}_z\right)^2. \tag{106}$$

It is interesting to point out that there is a certain similarity between this test and the one developed in [22] for universal decoding of memoryless Gaussian channels with an unknown deterministic interference. The decoding rule of [22] uses an auxiliary "backward channel" to maximize the empirical conditional entropy of the channel input given the channel output. In the case of only one unknown fading parameter, it is essentially equivalent to the test that minimizes $\inf_b \frac{1}{n}\sum_{t=1}^{n}\left(x_t^i - by_t\right)^2$. It should be emphasized, however, that this test is universal in the random coding sense, i.e., it attains the same random coding error exponent as the optimal ML decoder, whereas our detector (106) is universal in a somewhat stronger sense: for every specific code it universally achieves the largest possible fraction of the optimal ML error exponent associated with this code. $\Diamond$

As a final remark we point out that the detectors developed in this subsection can easily be extended to the case of $M > 2$, provided that M is held fixed while $n \to \infty$. Unfortunately, the extension to the general case is not trivial, and hence our results can not be applied to the problem of universal decoding at coding rates $R > 0$, where M grows exponentially with n.

4 Conclusions

In this paper, we studied the problem of composite hypothesis testing in the Neyman-Pearson formulation. By softening the false-alarm constraint, we considered a wider class of decision rules than in the generalized Neyman-Pearson criterion. This modification led to construction of an efficient test that attains exponential decay of the false-alarm and mis-detection probabilities with optimal exponents. We further derived a single-letter expression for the best false-alarm rate that can be achieved by a universal test with exponentially vanishing mis-detection probability. This in turn enabled us to furnish conditions on the geometry of the problem, under which efficient decision rules exist. As an additional benefit of our approach, we developed a test statistics, which is based on the worst-case ratio between the relative entropy of the empirical measure w.r.t. the true underlying probability measure and an optimal exponent of the LRT, and showed its asymptotic optimality under the competitive minimax criterion proposed in [18] for the Bayesian setting of the composite hypothesis testing problem.

Unfortunately, our results rely heavily on the fact that an observation sequence can be described by a finite dimensional vector of sufficient statistics. Therefore, our approach is not directly applicable to classes of hidden Markov sources (HMS), which are frequently used in speech recognition applications, and to general FS channels. However, we hope that, similarly to [10], [11], replacing the relative entropy by the sum of the log-likelihood function and the LZ complexity, it is possible to extend Theorems 1 and 4 to this case.

Another fundamental limitation of our analysis techniques is associated with the assumption that the number of hypotheses does not grow exponentially with n. While this assumption holds in a variety of interesting applications, the important problem of universal decoding cannot be formalized in our framework.

Appendix I

Proof of Corollary 3. First, by definition of ξ^*, there exist a sequence of decision rules Ω^n and $\zeta_n \to 0$ such that

$$-\frac{1}{n}\log P_e(\Omega^n|\theta) \geq \xi^* E^*(\theta) - \zeta_n, \forall\theta \in \Theta. \tag{107}$$

Since $\log(\cdot)$ is a monotonic increasing function and $P_e(\Omega^n|\theta) = \frac{1}{2}P_{e_1}(\Omega^n|\theta_1) + \frac{1}{2}P_{e_2}(\Omega^n|\theta_2)$ (assuming equiprobable messages), we obtain that

$$-\frac{1}{n}\log P_{e_1}(\Omega^n|\theta_1) + \frac{1}{n} \geq -\frac{1}{n}\log P_e(\Omega^n|\theta) \geq \xi^* E^*(\theta) - \zeta_n, \forall\theta \in \Theta. \tag{108}$$

Now, similarly to the proof of Theorem 1, for any $\boldsymbol{y} \in \Omega_2^n$ and $\theta \in \Theta$, we have

$$\begin{aligned}
\exp_2\{-n[\xi^* E^*(\theta) - \zeta_n - 1/n]\} &\geq P_{e_1}(\Omega^n|\theta_1) && (109)\\
&\geq \sum_{T(Q_{\boldsymbol{y}})\subseteq\Omega_2^n} \exp_2\{-n[D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \epsilon_n]\} && (110)\\
&\geq \exp_2\{-n[D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \epsilon_n]\} && (111)\\
&= \exp_2\{-n[D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \gamma_n + \epsilon_n - \gamma_n]\}, && (112)
\end{aligned}$$

where $\epsilon_n = |A|\log(n+1)/n$. Choosing $\gamma_n \geq \zeta_n + \epsilon_n + \frac{1}{n}$, we conclude that for all $\boldsymbol{y} \in \Omega_2^n$,

$$D(Q_{\boldsymbol{y}}||P_{\theta_1}) + \gamma_n \geq \xi^* E^*(\theta), \quad \forall\theta \in \Theta. \tag{113}$$

It means that $\Lambda_1^n(\gamma_n) \subseteq \Omega_1^n$ and hence,

$$e_2(\Lambda(\gamma)|\theta_2) \geq e_2(\Omega|\theta_2) \geq \xi^* E^*(\theta), \quad \forall\theta \in \Theta. \tag{114}$$

We next turn to the first kind error exponent. As in the proof of the part

(a) of Theorem 1, for all $\theta \in \Theta$,

$$P_{e_1}(\Lambda^n(\gamma_n)|\theta_1) = \sum_{T(Q_{\boldsymbol{y}}) \subseteq \Lambda_2^n(\gamma_n)} |T(Q_{\boldsymbol{y}})| \cdot p_{\theta_1}(\boldsymbol{y}) \tag{115}$$

$$\leq \sum_{T(Q_{\boldsymbol{y}}) \subseteq \Lambda_2^n(\gamma_n)} \exp_2\left\{-nD(Q_{\boldsymbol{y}}||P_{\theta_1})\right\} \tag{116}$$

$$\leq \sum_{T(Q_{\boldsymbol{y}})) \subseteq \Lambda_2^n(\gamma_n)} \exp_2\left\{-n[\xi^* E^*(\theta) - \gamma_n]\right\} \tag{117}$$

$$\leq \exp_2\left\{-n[\xi^* E^*(\theta) - \gamma_n - \epsilon_n]\right\}, \tag{118}$$

and therefore,

$$e_1(\Lambda(\gamma)|\theta_1) \geq \xi^* E^*(\theta), \quad \forall \theta \in \Theta. \tag{119}$$

Finally, by combining (114) with (119),

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_e(\Lambda^n(\gamma_n)|\theta) = \min\Big\{e_1(\Lambda(\gamma)|\theta_1), e_2(\Lambda(\gamma)|\theta_2)\Big\} \tag{120}$$

$$\geq \xi^* E^*(\theta), \quad \forall \theta \in \Theta, \tag{121}$$

which completes the proof.

□

Appendix II

Optimality of orthogonal signals for BSC's. Let $\mathcal{W}$ be the class of all BSC's with crossover probability $0 < p < 1$. We assume also that the signal set consists of only two messages of length n, $\boldsymbol{x}^1 = (x_1^1, x_2^1, \ldots, x_n^1)$ and $\boldsymbol{x}^2 = (x_1^2, x_2^2, \ldots, x_n^2)$. Using the symmetry of this problem, we can write (94) in the following simple form:

$$\lambda_{max}(p,q) = \inf_{p'}\left[qD(p'||p) + (1-q)D(p'||1-p)\right], \tag{122}$$

where q is the relative number of the coordinates in which $\boldsymbol{x}^1$ equals $\boldsymbol{x}^2$ and $D(\alpha||\beta)$ is the relative entropy between two binary sources with probabilities α and β, respectively. First, note that $\lambda_{max}(p,q)$ is concave in q because it is defined as the pointwise infimum of the collection of concave (in fact, affine) functions. Secondly, it can be seen that, for every p, $\lambda_{max}(p,\cdot)$ is symmetrical around $q = 1/2$, i.e., $\lambda_{max}(p,q) = \lambda_{max}(p,1-q)$. Combining these two properties, we have that

$$\lambda_{max}(p,q) = \frac{1}{2}\lambda_{max}(p,q) + \frac{1}{2}\lambda_{max}(p,1-q) \leq \lambda_{max}(p,1/2), \tag{123}$$

for all p and q. In other words, the signal set that universally attains the maximal false-alarm rate has the following structure: half of the coordinates of $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ are identical and another half of the coordinates are different, meaning that $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ are orthogonal signals.

References

[1] F.L. Lehmann, *Testing statistical Hypotheses.* New York: Wiley,1959.

[2] H. van Trees, *Detection, Estimation and Modulation Theory*, part I, John Wiley & Sons, New York 1968.

[3] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369-400, 1965.

[4] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401-408, Mar. 1989.

[5] M. Gutman, "Hypotheses testing with partial statistics and universal data compression," Ph.D. dissertation, Technion - I.I.T., 1987.

[6] J. Ziv, "Compression, tests for randomness and estimating the statistical model of an infdividual sequence," Sequences, R. M. Capocelli, Ed. New-York: Springer-Verlag, 1990, pp.366-373.

[7] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, pp. 278-286, Mar. 1988.

[8] N. Merhav, M. Gutman and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1014-1019, Sept. 1989.

[9] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109-1114, Sept. 1989.

[10] J. Ziv and N. Merhav, "Estimating the number of states of a finite-state source," *IEEE Trans. Inform. Theory*, vol. 38, pp. 61-65, Jan. 1992.

[11] N. Merhav, "Universal detection of messages via finite-state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242-2246, Sept. 2000.

[12] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains,"*IEEE Trans. Inform. Theory*, vol. IT-31, no. 3, pp. 360-365, May 1985.

[13] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, pp. 285-290, Mar. 1991.

[14] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-39, no. 10, pp. 2157-2166, October 1991.

[15] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. SAP-1, no. 1, pp. 90-100, January 1993.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[17] O. Zeitouni, J. Ziv and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, pp. 1597-1602, Sept. 1992.

[18] M. Feder and N. Merhav, "Competitive-minimax composite hypothesis testing and its applications," submitted for publication, 1999.

[19] L. Finesso, C.-C. Liu and P. Narayan, " The optimal error exponent for Markov order estimation," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1488-1497, Sept. 1996.

[20] Y. Migdal-Steinberg, "Large deviations bounds and universal hypotheses testing," M.Sc. Thesis, Technion - I.I.T., 1991.

[21] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Second Ed., Springer-Verlag New York, Inc. New York 1998.

[22] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1261-1269, July 1993.

[23] D. Whalen, *Detection of Signals in Noise*, Academic Press, 1971.

[24] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 420-430, Mar. 2000.

[25] Y. Steinberg and N. Merhav, "Identification in the Presence of Side Information with Application to Watermarking," to appear in *IEEE Trans. Inform. Theory*.

[26] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453-460, July 1985.

[27] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1726-1745, Sept. 1998.

[28] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, no. 5, pp. 530-536, Sept. 1978.