Universal Composite Hypothesis Testing: A Competitve Minimax Approach

Meir Feder* Neri Merhav†

April 9, 2001

In Memory of Aaron Daniel Wyner

Abstract

A novel approach is presented for the long-standing problem of composite hypothesis testing. In composite hypothesis testing, unlike in simple hypothesis testing, the probability function of the observed data given the hypothesis, is uncertain as it depends on the unknown value of some parameter. The proposed approach is to minimize the worst–case ratio between the probability of error of a decision rule that is independent of the unknown parameters and the minimum probability of error attainable given the parameters. The principal solution to this minimax problem is presented and the resulting decision rule is discussed. Since the exact solution is, in general, hard to find, and a-fortiori hard to implement, an approximation method that yields an asymptotically minimax decision rule is proposed. Finally, a variety of potential application areas are provided in signal processing and communications with special emphasis on universal decoding.

Index Terms: Composite hypothesis testing, universal decoding, error exponents, maximum likelihood, likelihood ratio, generalized likelihood ratio test.

*M. Feder is with the Department of Electrical Engineering – Systems, Tel Aviv University, Tel Aviv, 69978, Israel. E-mail: `meir@eng.tau.ac.il`

†N. Merhav is with the Department of Electrical Engineering, Technion – I.I.T., Haifa 32000, Israel. E-mail: `merhav@ee.technion.ac.il` .

1 Introduction

Composite hypothesis testing is a long–standing problem in statistical inference which still lacks a satisfactory solution in general. In composite hypothesis testing (see, e.g., [19, Sect. 9.3], [27, Sect. 2.5]) the problem is to design a test, or a decision rule, for deciding in favor of one out of several hypotheses, under some uncertainty in the parameters of the probability distribution (or density) functions associated with these hypotheses. This uncertainty precludes the use of the optimal likelihood ratio test (LRT) or the maximum-likelihood (ML) decision rule.

Composite hypothesis testing finds its applications in a variety of problem areas in signal processing and communications where the aforementioned uncertainty exists in some way. A few important examples are: (i) signal detection in the presence of noise, where certain parameters of the desired signal (e.g., amplitude, phase, Doppler shift) are unknown [9], [28], (ii) pattern recognition problems like speech recognition [20] and optical character recognition [25], (iii) model order selection [1],[21], for instance, estimating the order of a Markov process [16], and (iv) universal decoding in the presence of channel uncertainty [2, Chap. 2, Sect. 5], [5],[11],[14],[30]. The latter application, which will receive special attention in this paper, is actually the one that motivated our general approach in the first place.

We begin with an informal description of the problem and the general approach proposed in this paper. To fix ideas, let us consider the binary case, i.e., the case where there are only two hypotheses H_i, $i = 0, 1$. Of course, the discussion below and the main results described will extend to multiple hypotheses, and so, this restriction to binary hypothesis testing is merely for the sake of simplicity of the exposition. As mentioned above, in composite hypothesis testing the probability function of the observed data given either hypothesis depends on the unknown value of a certain index, or parameter. Specifically, for each hypothesis H_i, $i = 0, 1$, there is a family of probability density functions (PDF's) $\{p_{\theta_i}(\boldsymbol{y}|H_i),\ \theta_i \in \Lambda_i\}$,[1] where $\boldsymbol{y} = (y_1, ..., y_n)$ is a sequence of observations taking on values in the observation space $\mathcal{Y}^n$, θ_i is the index of the PDF within the family (most commonly, but not necessarily, θ_i is a parameter vector of a smooth parametric family), and Λ_i is the

[1] The duplication of the index i is meant to cover a general case where each hypothesis may be associated with its own parameter(s). There are, however, cases (cf. Section 4.2) where the same parameter(s) determine the distribution under all hypotheses.

index set, $i = 0, 1$.

A decision rule, or a test Ω is sought, ideally, to minimize $P_e(\Omega|\theta_0, \theta_1)$, the probability of error associated with Ω and induced by the true values of θ_0 and θ_1 under both hypotheses, where for the sake of simplicity, it will be assumed that the two hypotheses are a–priori equiprobable. As is well–known, the optimal test for simple hypotheses (i.e., known θ_0 and θ_1) is the ML test, or the LRT, denoted $\Omega^*(\theta_0, \theta_1)$, which is based on comparing the likelihood ratio

$$L(\boldsymbol{y}) = \frac{p_{\theta_1}(\boldsymbol{y}|H_1)}{p_{\theta_0}(\boldsymbol{y}|H_0)} \tag{1}$$

to a certain threshold (whose value is one in the case of a uniform prior). The minimum error probability associated with $\Omega^*(\theta_0, \theta_1)$ will be denoted by $P_e^*(\theta_0, \theta_1)$.

In general, the optimum LRT becomes inapplicable in the lack of exact knowledge of θ_0 and θ_1 unless it happens to be independent of those parameters, namely, a uniformly most powerful test exists. In other situations, there are two classical approaches to composite hypothesis testing. The first is a Bayesian approach, corresponding to an assumption of a certain prior $\mu(\theta_i|H_i)$ on Λ_i for each hypothesis. This assumption converts the composite hypothesis problem into a simple hypothesis testing problem w.r.t. the mixture densities $\tilde{p}(\boldsymbol{y}|H_i) = \int_{\Lambda_i} p_{\theta_i}(\boldsymbol{y}|H_i)\mu(\theta_i|H_i)d\theta_i$, $i = 0, 1$, and is hence optimally solved (in the sense of the expected error probability w.r.t. θ_0 and θ_1) by the LRT w.r.t. those densities. Unfortunately, the Bayesian approach suffers from several weaknesses: First, the assumption that the prior $\mu(\cdot|H_i)$ is known, not to mention the assumption that it at all exists, is hard to justify in most applications. Secondly, even if existent and known, the averaging w.r.t. this prior is not very appealing because once θ_i is drawn, it remains fixed throughout the entire experiment. Finally, on the practical side, the above–defined mixture PDF's $\tilde{p}(\boldsymbol{y}|H_i)$ are hard to compute in general.

The second approach, which is most commonly used, is the generalized likelihood ratio test (GLRT) [27, p. 92]. In the GLRT approach, the idea is to implement an LRT with the unknown θ_i being replaced by their ML estimates under the two hypotheses. More precisely, the GLRT compares the generalized likelihood ratio

$$\frac{\sup_{\theta_1 \in \Lambda_1} p_{\theta_1}(\boldsymbol{y}|H_1)}{\sup_{\theta_0 \in \Lambda_0} p_{\theta_0}(\boldsymbol{y}|H_0)} \tag{2}$$

to a suitable threshold. Although in some situations the GLRT is asymptotically optimum in a certain sense (see, e.g., [29] for necessary and sufficient conditions in a Neyman–Pearson like setting, [17] for asymptotic minimaxity, and [2, p. 165, Theorem 5.2] for universal decoding over discrete memoryless channels), it still lacks a solid theoretical justification in general. Indeed, there are examples where the GLRT is strictly suboptimum even asymptotically. One, rather synthetic, example can be found in [11, Sect. III, pp. 1754–1755]. In another, perhaps more natural, example associated with the additive Gaussian channel (see Appendix), it is shown that the GLRT is uniformly worse than another decision rule that is independent of θ. Moreover, in some situations, the GRLT becomes altogether totally useless. For example, if the two classes are nested, that is, if $\Lambda_0 \subset \Lambda_1$ and $p_{\theta_i}(\cdot|H_i)$ depends on the hypothesis H_i only via θ_i $(i = 0, 1)$, then the generalized likelihood ratio (2) can never be less than unity, and so, H_1 would always be preferred (unless, of course, the threshold is larger than unity).

In this paper, we propose a new approach to composite hypothesis testing. According to this approach, we seek a decision rule that is independent of the unknown θ_0 and θ_1, and whose performance is nevertheless uniformly as close as possible to that of the optimum LRT $\Omega^*(\theta_0, \theta_1)$ for all $(\theta_0, \theta_1) \in \Lambda_0 \times \Lambda_1$. More precisely, we seek an optimum decision rule Ω in the sense of the *competitive minimax*:

$$K_n \stackrel{\triangle}{=} \min_{\Omega} \max_{(\theta_0,\theta_1)\in\Lambda_0\times\Lambda_1} \frac{P_e(\Omega|\theta_0, \theta_1)}{P_e^*(\theta_0, \theta_1)}. \tag{3}$$

The ratio $P_e(\Omega|\theta_0, \theta_1)/P_e^*(\theta_0, \theta_1)$ designates the loss incurred by employing a decision rule Ω that is ignorant of (θ_0, θ_1), relative to the optimum LRT for that (θ_0, θ_1). To make this loss uniformly as small as possible across $\Lambda_0 \times \Lambda_1$, we seek a decision rule that minimizes the worst–case value of this ratio, i.e., its maximum. This idea of competitive (or, relative) minimax with respect to optimum performance for known (θ_0, θ_1), has the merit of partially compensating for the inherently pessimistic nature of the minimax criterion.

As a general concept, the competitive minimax criterion is by no means new. For example, the very same approach has been used to define the notion of the minimax redundancy in universal source coding [3], where a coding scheme is sought that minimizes the worst–case loss of coding length beyond the entropy of the source. Moreover, even within the framework of composite hypothesis testing, two ideas in the same spirit have been studied in the Neyman–Pearson setting of the problem, although in a substantially different manner: The first, referred to as the *exponential rate optimal* (ERO) test, was proposed first by Hoeffding [10], extended later by Tusnády [26], and further developed in the information theory literature by Ziv [31], Gutman [8] and others. In this series of works, it is demonstrated that there exist tests that maximize the error exponent of the second kind, uniformly across all alternatives, subject to a uniform constraint on the error exponent of the first kind across all probability measures corresponding to the null hypothesis. The shortcoming of the ERO approach, however, is that there may always exist probability measures corresponding to the alternative hypothesis, for which the probability of error of the second kind tends to unity.[2] The second idea, in this spirit of a competitive minimax criterion, is the notion of a *most stringent test* [12, pp. 339–341], where the minimax is taken on the difference, rather than the ratio, between the powers of the two tests.

The advantage of addressing the ratio between probabilities as proposed herein (3) is that it corresponds to the difference between the exponential rates of the error probabilities: As is well-known, under most commonly–used probabilistic models (e.g., i.i.d. and Markov sources/channels), $P_e^*(\theta_0, \theta_1)$ normally decays exponentially rapidly as a function of n, the dimension of the observed data set $\boldsymbol{y}$. Thus, if the value of K_n happens to be a subexponential function of n, i.e., $\lim_{n\to\infty} n^{-1} \ln K_n = 0$, this means that, uniformly over $\Lambda_0 \times \Lambda_1$, the exponential rate of $P_e(\Omega|\theta_0, \theta_1)$, for the Ω that attains (3), is as good as that of the optimum LRT for known (θ_0, θ_1). In this case, Ω is said to be a *universal decision rule in the error exponent sense.*

The exact solution to the competitive minimax problem is, in general, hard to find, and a-fortiori, hard to implement. Fortunately enough, it turns out that these difficulties are at least partially alleviated if one is willing to resort to suboptimal solutions that are asymptotically optimal. The key observation that opens the door in this direction is that in order for a decision rule to be universal in the error exponent sense defined above, it need not be strictly minimax, but may only be asymptotically minimax in the sense that it achieves (3) within a factor that grows subexponentially with n. A major goal of the paper is to develop and investigate such asymptotically minimax decision rules.

The outline of the paper is as follows. In Section 2, we first characterize and analyze the structure of the competitive minimax decision rule. We will also obtain expressions for the minimax value K_n, and thereby furnish conditions for the existence of a universal decision rule in the error exponent sense. As mentioned earlier, the strictly competitive

[2] In a recent paper [13], the competitive minimax approach considered here is combined with the ERO approach and this difficulty is alleviated by allowing the constraint on the first error exponent to depend on the (unknown) probability measure of the null hypothesis.

minimax–optimal decision rule in the above–defined sense is hard to derive in general. In Section 3, we present several approximate decision rules that yield asymptotically the same (or almost the same) error exponent as this decision rule, but with the advantage of having explicit forms and performance evaluation in many important practical cases. In Section 4, we present applications of universal hypothesis testing in certain communications and signal processing problems, and elaborate on the universal decoding problem in communication via unknown channels. Finally, in Section 5, we conclude by listing some open problems.

2 The Competitive Minimax Criterion

In this section, we provide a precise formulation of the competitive minimax approach for multiple composite hypothesis testing, and then study the structure and general properties of the minimax–optimal decision rule.

Let $\boldsymbol{y} = (y_1, ..., y_n)$ denote an n-dimensional vector of observations, where each coordinate y_i, $i = 1, \ldots, n$, takes on values in a certain alphabet $\mathcal{Y}$ (e.g., a finite alphabet, a countable alphabet, an interval, or the entire real line). The n-th order Cartesian power of $\mathcal{Y}$, which is the space of n–sequences, will be denoted by $\mathcal{Y}^n$. There are M ($M \geq 2$ – integer) composite hypotheses, $H_0, \ldots, H_{M-1}$, regarding the probabilistic information source that has generated $\boldsymbol{y}$. Associated with each hypothesis H_i, $i = 0, \ldots, M-1$, there is a family of probability measures on $\mathcal{Y}^n$ that possess jointly measurable Radon–Nykodim derivatives (w.r.t. a common dominating measure[3]), $\{p_{\theta_i}(\boldsymbol{y}|H_i),\ \theta_i \in \Lambda_i\}$, where θ_i is the parameter, or more generally, the index of the probability measure within the family and Λ_i is the index set, $i = 0, \ldots, M-1$. For convenience,[4] will also denote $\theta = (\theta_0, \ldots, \theta_{M-1})$ and $\Lambda = \Lambda_0 \times \ldots \times \Lambda_{M-1}$. In some situations of practical interest θ may not be free to take on values across the whole Cartesian product $\Lambda_0 \times \ldots \times \Lambda_{M-1}$ but only within a certain subset as the components $\theta_0, \theta_1, \ldots, \theta_{M-1}$ may be related to each other (see, e.g., subsection 4.2, where even $\theta_0 = \theta_1 = \ldots = \theta_{M-1}$). In such cases, it will be understood that Λ stands for the set of allowable combinations of $(\theta_0, \theta_1, \ldots, \theta_{M-1})$.

A decision rule is a (possibly randomized) map $\Omega : \mathcal{Y}^n \to \{0, \ldots, M-1\}$, characterized

[3]The dominating measure will be assumed the counting measure in the discrete case, or the Lebesgue measure in the continuous case.

[4]This is meant to avoid cumbersome notation when denoting quantities that depend on $\theta_0, \ldots, \theta_{M-1}$, such as the probability of error.

by a conditional probability vector function

$$\Omega = \{(\omega(0|\boldsymbol{y}), \ldots, \omega(M-1|\boldsymbol{y})),\ \boldsymbol{y} \in \mathcal{Y}^n\},$$

with $\omega(i|\boldsymbol{y})$ being the conditional probability of deciding in favor of H_i given $\boldsymbol{y}$, $i = 0, \ldots, M-1$. Of course, $\omega(i|\boldsymbol{y})$ is never negative and $\sum_{i=0}^{M-1} \omega(i|\boldsymbol{y}) = 1$ for all $\boldsymbol{y} \in \mathcal{Y}^n$. If a test is deterministic, then for every $\boldsymbol{y}$ and i, $\omega(i|\boldsymbol{y})$ is either zero or one, in which case Ω_i will designate the subset of $\boldsymbol{y}$-vectors for which $\omega(i|\boldsymbol{y}) = 1$, $i = 0, \ldots, M-1$. For a given decision rule Ω and $0 \leq i \leq M-1$, let

$$\begin{aligned} P_e(\Omega|\theta_i) &= \int_{\mathcal{Y}^n} [1 - \omega(i|\boldsymbol{y})] p_{\theta_i}(\boldsymbol{y}|H_i) d\boldsymbol{y} \\ &= \int_{\Omega_i^c} p_{\theta_i}(\boldsymbol{y}|H_i) d\boldsymbol{y} \quad \text{for a deterministic decision rule.} \end{aligned} \tag{4}$$

The (overall) probability of error, for a uniform prior on $\{H_i\}$, is given by

$$P_e(\Omega|\theta) = \frac{1}{M} \sum_{i=0}^{M-1} P_e(\Omega|\theta_i). \tag{5}$$

Let $\Omega^*(\theta) = \{\Omega_0^*(\theta), \ldots, \Omega_{M-1}^*(\theta)\}$ denote the optimum ML decision rule, i.e.,

$$\Omega_i^*(\theta) = \{\boldsymbol{y} :\ p_{\theta_i}(\boldsymbol{y}|H_i) \geq \max_{j \neq i} p_{\theta_j}(\boldsymbol{y}|H_j)\}, \quad i = 0, \ldots, M-1, \tag{6}$$

where ties are broken arbitrarily, and denote

$$P_e^*(\theta) = P_e(\Omega^*(\theta)|\theta). \tag{7}$$

Finally, define

$$K_n(\Omega, \theta) = \frac{P_e(\Omega|\theta)}{P_e^*(\theta)} \tag{8}$$

and the competitive minimax is defined as

$$K_n = \inf_{\Omega} \sup_{\theta \in \Lambda} K_n(\Omega, \theta). \tag{9}$$

While in simple hypothesis testing, the optimal ML decision rule $\Omega^*(\theta)$ is clearly deterministic, it turns out that in the composite case, the competitive minimax criterion considered here, may yield a randomized decision rule as an optimum solution. Intuitively, this randomization gives rise to a certain compromise among the different ML decision rules corresponding to different values of θ. The competitive minimax criterion defined in eq. (9) is equivalent to

$$\inf_{\Omega} \sup_{\theta \in \Lambda} \frac{\frac{1}{M} \sum_{i=0}^{M-1} \int_{\mathcal{Y}^n} [1 - \omega(i|\boldsymbol{y})] p_{\theta_i}(\boldsymbol{y}|H_i) d\boldsymbol{y}}{P_e^*(\theta)}. \tag{10}$$

A common method to solve the minimax problem (10) is to use a "mixed strategy" for θ. Specifically, note that eq. (10) can be written as

$$\begin{aligned} K_n &= \inf_{\Omega}\sup_{\mu}\int_{\Lambda}\frac{\mu(d\theta)}{P_e^*(\theta)}\cdot\frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]p_{\theta_i}(\boldsymbol{y}|H_i)d\boldsymbol{y} \\ &\triangleq \inf_{\Omega}\sup_{\mu}\bar{K}_n(\Omega,\mu) \end{aligned} \tag{11}$$

where $\mu(\cdot)$ is a probability measure on Λ (defined on a suitably chosen sigma–algebra of Λ). Note that both μ and Ω range over convex sets (as both are probability measures) and that $\bar{K}_n$ is a convex–concave functional (in fact, affine in both arguments). Therefore, if $\{p_{\theta_i}(\cdot|H_i),\ \theta_i\in\Lambda_i\}$ are such that: (i) the space of decision rules $\{\Omega\}$ is compact, and (ii) $\bar{K}_n(\cdot,\mu)$ is continuous for every μ (which is obviously the case, for example, when $|\mathcal{Y}|<\infty$), then the minimax value is equal to the maximin value [24, Theorem 4.2], i.e.,

$$K_n = \sup_{\mu}\inf_{\Omega}\int_{\Lambda}\frac{\mu(d\theta)}{P_e^*(\theta)}\cdot\frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]p_{\theta_i}(\boldsymbol{y}|H_i)d\boldsymbol{y}. \tag{12}$$

For a given μ, the minimizer Ω of $\bar{K}_n(\Omega,\mu)$ is clearly given as follows. Let

$$f_i(\boldsymbol{y}) = \int_{\Lambda}\frac{\mu(d\theta)p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}. \tag{13}$$

Then, $\omega(i|\boldsymbol{y})$ is given by

$$\omega(i|\boldsymbol{y}) = \begin{cases} 1 & \text{if } f_i(\boldsymbol{y}) > \max_{j\neq i} f_j(\boldsymbol{y}) \\ 0 & \text{if } f_i(\boldsymbol{y}) < \max_{j\neq i} f_j(\boldsymbol{y}) \\ \text{arbitrary value in } [0,1] & \text{if } f_i(\boldsymbol{y}) = \max_{j\neq i} f_j(\boldsymbol{y}) \end{cases}. \tag{14}$$

The last line in the above equation tells us that any probability distribution $\omega(\cdot|\boldsymbol{y})$ over the set of indices $\{i\}$ that maximize $f_i(\boldsymbol{y})$ is a solution to the inner minimization problem of eq. (12). The maximizing weight function $\mu^*(\cdot)$ (whenever exists) can be found by substituting the solution (14) into (12) and maximizing the resulting expression over μ. The resulting expression is therefore:

$$\sup_{\mu}\frac{1}{M}\int_{\mathcal{Y}^n}[\sum_{i=1}^{M} f_i(\boldsymbol{y}) - \max_i f_i(\boldsymbol{y})]d\boldsymbol{y} \tag{15}$$

Note that eq. (15) is also the minimax value of (11), since the minimax and maximin values coincide. This does not imply, however, that any maximin decision rule is necessarily minimax. Nonetheless, whenever there exists a saddle–point (Ω^*,μ^*) it is both minimax and maximin. In this case, the desired minimax decision rule is of the form of eq. (14), but with $\mu=\mu^*$ and with certain values of $\omega(i|\boldsymbol{y})\in(0,1)$ for randomized tie–breaking. We next demonstrate the intricacy of this problem by example.

Example. Consider a binary symmetric channel (BSC) whose unknown crossover probability θ can either take the value $0 \le \alpha < 1/2$ or the value $1/2 < \beta \le 1$, where α and β are given and known. Let a single bit $x \in \{0,1\}$ be transmitted across the channel and let $y \in \{0,1\}$ be the observed channel output. The problem of decoding x upon observing y under the uncertainty of whether $\theta = \alpha$ or $\theta = \beta$ is, of course, a problem of binary composite hypothesis testing, where according to hypothesis H_0, $x = 0$ was transmitted, and according to H_1, $x = 1$. In this case, we have

$$\begin{aligned} p_\theta(y=0|H_0) &= 1 - p_\theta(y=1|H_0) &= 1-\theta \\ p_\theta(y=0|H_1) &= 1 - p_\theta(y=1|H_1) &= \theta \end{aligned} \tag{16}$$

The ML decoder for $\theta = \alpha$ accepts H_0 for $y = 0$ and H_1 for $y = 1$, whereas for $\theta = \beta$ it makes the opposite decisions. The resulting error probabilities are, therefore, $P_e^*(\alpha) = \alpha$ and $P_e^*(\beta) = 1 - \beta$. To describe the minimax decoder, we have to specify the weights assigned to α and β. Let $\mu = \mu(\alpha) = 1 - \mu(\beta)$, and for a given value of μ, let

$$\begin{aligned} A_\mu &\triangleq \frac{\mu}{\alpha}\cdot(1-\alpha) + \frac{1-\mu}{1-\beta}\cdot(1-\beta) \\ &= \frac{\mu}{\alpha}\cdot(1-\alpha) + 1 - \mu, \end{aligned} \tag{17}$$

and

$$\begin{aligned} B_\mu &\triangleq \frac{\mu}{\alpha}\cdot\alpha + \frac{1-\mu}{1-\beta}\cdot\beta \\ &= \mu + \frac{1-\mu}{1-\beta}\cdot\beta. \end{aligned} \tag{18}$$

Denoting $\omega_0 = \omega(0|0)$, $\omega_1 = \omega(0|1)$, and $\Omega = (\omega_0, \omega_1)$, and setting $n = 1$, we now have

$$\begin{aligned} \bar{K}_1(\Omega,\mu) &= \frac{\mu}{\alpha}\{\frac{1}{2}[(1-\alpha)(1-\omega_0) + \alpha(1-\omega_1)] + \frac{1}{2}[\alpha\omega_0 + (1-\alpha)\omega_1]\} + \\ &\quad \frac{1-\mu}{1-\beta}\{\frac{1}{2}[(1-\beta)(1-\omega_0) + \beta(1-\omega_1)] + \frac{1}{2}[\beta\omega_0 + (1-\beta)\omega_1]\} \\ &= \frac{1}{2}[A_\mu + B_\mu + (\omega_1 - \omega_0)(A_\mu - B_\mu)] \\ &\triangleq \frac{1}{2}[A_\mu + B_\mu + \Delta(A_\mu - B_\mu)] \end{aligned} \tag{19}$$

where the last two lines tell us that the performance of the decoder (in the competitive minimax sense) depends on ω_0 and ω_1 only via the difference Δ between them, and so, with a slight abuse of notation, we will denote the last line of eq. (19) by $\bar{K}_1(\Delta,\mu)$. If we can find a saddle–point (Δ^*, μ^*) of $\bar{K}_1(\Delta,\mu)$, then the decision rule Ω^* corresponding to Δ^* would be a minimax–optimal. As is well known [22, Lemma 36.2], the pair (Δ^*, μ^*) where Δ^*

minimizes $\max_\mu \bar{K}_1(\Delta,\mu)$ and where μ^* maximizes $\min_\Delta \bar{K}_1(\Delta,\mu)$, is such a saddle–point of $\bar{K}_1$. Now, the maximin decision rule (14) estimates x by $\hat{x}$ using the following rules:
If $A_\mu > B_\mu$, then $\hat{x} = y$.
If $A_\mu < B_\mu$, then $\hat{x} = 1 - y$.
If $A_\mu = B_\mu$,

$$\hat{x} = \begin{cases} 0 & \text{with probability } \omega_y \\ 1 & \text{with probability } 1-\omega_y \end{cases} \tag{20}$$

It then follows (as can also be seen directly from the expression of $\bar{K}_1(\Delta,\mu)$) that the performance of this decision rule for a given μ is given by

$$\min_\Delta \bar{K}_1(\Delta,\mu) = \min\{A_\mu, B_\mu\}.$$

The maximum of this expression w.r.t. μ occurs when $A_\mu = B_\mu$, (corresponding in turn to the above–described randomized mode of the decoder), which is achieved for

$$\mu = \mu^* \stackrel{\triangle}{=} \frac{\alpha(2\beta-1)}{(1-2\alpha)(1-\beta)+\alpha(2\beta-1)}, \tag{21}$$

and so, the maximin value (which is also the minimax value) is given by

$$K_1 = \max_\mu \min_\Delta \bar{K}_1(\Delta,\mu) = \frac{\beta-\alpha}{(1-2\alpha)(1-\beta)+\alpha(2\beta-1)}.$$

Solving now the minimax problem of $K_1(\Delta,\mu)$, we obtain after some standard algebraic manipulations

$$\Delta^* = \frac{\alpha+\beta-1}{(1-2\alpha)(1-\beta)+\alpha(2\beta-1)},$$

which is always in $[-1,1]$ and hence can be realized as a difference between some two numbers ω_1^* and ω_0^* in $[0,1]$. Thus, unless Δ^* happens to be equal to 1, 0, or -1, the minimax decoder must be randomized. $\diamond$

This example is interesting, not only in that the minimax decoder is randomized, but also because the weight function $\mu(\cdot)$ is such that the test statistic $f_i(\boldsymbol{y})$ (cf. eq. (13)) has no unique maximum. It turns out that as n grows, and as the index sets Λ_i become more complicated, the test statistic $f_i(\boldsymbol{y})$ gives rise to a larger degree of discrimination among the hypotheses, the need for randomization reduces, and the weight function $\mu(\cdot)$ has a weaker effect on the decision rule and its performance. Furthermore, it becomes increasingly more difficult to devise the exact minimax decision rule in closed form. Fortunately, as will be seen in the next section, one can approximate $f_i(\boldsymbol{y})$ and the resulting (deterministic) decision rule turns out to be asymptotically minimax under fairly mild regularity conditions.

We conclude this section by further characterization of the value of the minimax-maximin game,

$$K_n = \inf_{\Omega} \sup_{\theta \in \Lambda} \frac{P_e(\Omega|\theta)}{P_e^*(\theta)} = \sup_{\mu} \inf_{\Omega} \int_{\Lambda} \mu(d\theta) \frac{P_e(\Omega|\theta)}{P_e^*(\theta)}. \tag{22}$$

To make the derivation simpler, we begin with the case of two hypotheses and assume that Λ is a finite set. By plugging the optimum (Bayesian) decoder for a given μ, we have

$$K_n = \frac{1}{2} \sup_{\mu} \int_{\mathcal{Y}^n} d\boldsymbol{y} \min_i \sum_{\theta \in \Lambda} \mu(\theta) \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}, \tag{23}$$

where we note that $P_e^*(\theta) = \frac{1}{2} \int_{\mathcal{Y}^n} d\boldsymbol{y} \min_i p_\theta(\boldsymbol{y}|H_i)$, thus for every μ, we have

$$\frac{1}{2} \int_{\mathcal{Y}^n} d\boldsymbol{y} \sum_{\theta \in \Lambda} \mu(\theta) \min_i \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)} = 1. \tag{24}$$

In view of this, the factor K_n can be thought of as arising from interchanging the order between the minimization over i and the summation over Λ. Since K_n can also be thought of as the ratio between the expressions of eqs. (23) and (24), we now further examine this ratio. We start with the l.h.s. of eq. (24), which is the denominator of this ratio. Since eq. (24) holds for any μ, we may select μ to be the uniform distribution and then

$$\frac{1}{2} \int_{\mathcal{Y}^n} d\boldsymbol{y} \sum_{\theta \in \Lambda} \mu(\theta) \min_i \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)} \geq \frac{1}{2|\Lambda|} \int_{\mathcal{Y}^n} d\boldsymbol{y} \max_{\Lambda} \min_i \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}. \tag{25}$$

On the other hand, the r.h.s. of eq. (23) is upper bounded by

$$\frac{1}{2} \sup_{\mu} \int_{\mathcal{Y}^n} d\boldsymbol{y} \min_i \sum_{\theta \in \Lambda} \mu(\theta) \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)} \leq \frac{1}{2} \int_{\mathcal{Y}^n} d\boldsymbol{y} \min_i \max_{\Lambda} \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}. \tag{26}$$

Combining eqs. (23), (24), (25), and (26), we get

$$K_n \leq |\Lambda| \frac{\int_{\mathcal{Y}^n} d\boldsymbol{y} \min_i \max_{\Lambda} [p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)]}{\int_{\mathcal{Y}^n} d\boldsymbol{y} \max_{\Lambda} \min_i [p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)]}. \tag{27}$$

As can be seen, there are two factors on the r.h.s. The first is the size of index set Λ, which accounts for its richness, and measures the degree of a–priori uncertainty regarding the true value of the index or the parameter. The second factor is a ratio between two expressions which depends more intimately on the structure and the geometry of the problem. Accordingly, a sufficient condition for the existence of universal decision rules refers both to the richness of the class and its structure. Note, in particular, that if the minimax at the integrand of the numerator of eq. (27) happens to agree with the maximin at the denominator for every $\boldsymbol{y}$, (which is the case in certain examples), then $K_n \leq |\Lambda|$.

In the more general case of M hypotheses, let us define the following operator over a function whose argument takes M values.

$$\text{Min}_i f(i) = \sum_{i=1}^{M} f(i) - \max_i f(i) \tag{28}$$

In other words, $\text{Min}_i f(i)$ is the sum of all terms except for the maximal term of $f(i)$. We then have that K_n is upper bounded by the same expression as in eq. (27) except that the ordinary minimum over i is replaced by Min over i.

In certain examples, we can analyze this expression and determine whether it behaves subexponentially with n, in which case, a universal decision rule exists in the error exponent sense. As is well–known, and will be discussed in Section 4, for the problem of decoding a randomly chosen block code, in the presence of an unknown channel from a sufficiently regular class, there exist universal decision rules (universal decoders) in the error exponent sense.

3 Approximations and Suboptimal Decision Rules

The decoder developed in the previous section is hard to implement, in general, for the following reasons. First, the minimax decoder that attains (10) and has the structure given by eqs. (13) - (14), is not given explicitly as it depends on the least favorable weight function $\mu^*(\cdot)$, which is normally hard to find. Secondly, an exact closed–form expression of $P_e^*(\theta)$, which is necessary for explicit specification of the decision rule, is rarely available. Finally, even if both $\mu^*(\cdot)$ and $P_e^*(\theta)$ are given explicitly, the mixture integral of eq. (13) is prohibitively complicated to calculate in most cases.

In this section, we propose two strategies of controlling the compromise between performance and ease of implementation. The first (subsection 3.1.) leads to asymptotically optimal performance (in the competitive minimax sense) under certain conditions. The second strategy (subsections 3.2,3.3) might be suboptimal, yet it is easy to characterize its guaranteed performance.

3.1 An Asymptotically Minimax Decision Rule

In this subsection, we approximate the minimax decision rule by a decision rule $\hat{\Omega}$, which is, on the one hand, easier to implement, and on the other hand, under fairly mild regularity conditions, *asymptotically minimax*, i.e.,

$$\sup_{\theta\in\Lambda}\frac{P_e(\hat{\Omega}|\theta)}{P_e^*(\theta)} \le L_n\cdot\inf_{\Omega}\sup_{\theta\in\Lambda}\frac{P_e(\Omega|\theta)}{P_e^*(\theta)} = L_nK_n, \tag{29}$$

where the sequence $\{L_n\}$ grows subexponentially in n, i.e., $\lim_{n\to\infty}\frac{1}{n}\ln L_n = 0$. Note that if, in addition, K_n is subexponential as well, then so is the product L_nK_n, and then $P_e^*(\hat{\Omega}|\theta)$ is of the same exponential rate (as a function of n) as $P_e^*(\theta)$ for every θ uniformly in Λ.

Consider the test statistic

$$\hat{f}_i(\boldsymbol{y}) = \sup_{\theta\in\Lambda}\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}, \quad i=0,\ldots,M-1, \tag{30}$$

and let the decision rule $\hat{\Omega} = (\hat{\Omega}_0,\ldots,\hat{\Omega}_{M-1})$ be defined by

$$\hat{\Omega}_i = \{\boldsymbol{y}:\ \hat{f}_i(\boldsymbol{y}) \ge \max_{j\ne i}\hat{f}_j(\boldsymbol{y})\}, \quad i=0,\ldots,M-1 \tag{31}$$

where ties are broken arbitrarily. Observe that this is a variant of the GLRT except that, prior to the maximization over Λ, the likelihood functions corresponding to the different hypotheses, are first normalized by $P_e^*(\theta)$, thus giving higher weights to parameter values for which the hypotheses are more easily distinguishable (i.e., where $P_e^*(\theta)$ is relatively small). Intuitively, this manifests the fact that this decision rule strives to capture the relatively good performance of the ML decision rule at these points.

We next establish the asymptotic minimaxity of $\hat{\Omega}$. To this end, let us define the following two functionals:

$$K_n(\Omega) = \sup_{\theta\in\Lambda}\frac{P_e(\Omega|\theta)}{P_e^*(\theta)} = \sup_{\theta\in\Lambda}\frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}d\boldsymbol{y} \tag{32}$$

and

$$\hat{K}_n(\Omega) = \frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\sup_{\theta\in\Lambda}\left[\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}\right]d\boldsymbol{y} = \frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\hat{f}_i(\boldsymbol{y})d\boldsymbol{y}. \tag{33}$$

Note that the expression of $\hat{K}_n(\Omega)$ is similar to that of $K_n(\Omega)$ except that the supremum over Λ is interchanged with the integration and summation. Therefore, $K_n(\Omega) \le \hat{K}_n(\Omega)$ for every Ω. Note also that while the minimax decision rule minimizes $K_n(\Omega)$, the decision rule $\hat{\Omega}$ minimizes $\hat{K}_n(\cdot)$. The following Theorem uses these facts to give an upper bound to the performance of $\hat{\Omega}$ (in the competitive minimax sense) in terms of the optimal value K_n.

Theorem 1 *Let $\hat{\Omega}$ be defined as in eq. (31) and let*

$$L_n \stackrel{\triangle}{=} \sup_{\Omega} \frac{\hat{K}_n(\Omega)}{K_n(\Omega)}.$$

Then,

$$K_n(\hat{\Omega}) \leq L_n K_n.$$

Proof. Combining the two facts mentioned in the paragraph that precedes Theorem 1, we have, for every decision rule Ω,

$$K_n(\hat{\Omega}) \leq \hat{K}_n(\hat{\Omega}) \leq \hat{K}_n(\Omega) \leq L_n K_n(\Omega), \tag{34}$$

and the proof is completed by minimizing the right–most side w.r.t. Ω. $\square$

In view of the foregoing discussion on asymptotic minimaxity, Theorem 1 is especially interesting in cases where the sequence $\{L_n\}$ happens to be subexponential. As we shall see next in a few examples, this is the case as long as the families of sources, corresponding to the different hypotheses, are not too "rich." While the exact value of L_n might be difficult to compute in general, its subexponential behavior can still be established by upper bounds.

Examples

1. Finite index sets. Suppose that Λ_i, $i = 0, ..., M-1$, are all finite sets and let $L \stackrel{\triangle}{=} |\Lambda| = \prod_{i=0}^{M-1} |\Lambda_i|$. Then, for every Ω,

$$\begin{aligned}
\hat{K}_n(\Omega) &= \frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\max_{\theta}\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}d\boldsymbol{y} \\
&\leq \frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\sum_{\theta=1}^{L}\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}d\boldsymbol{y} \\
&= \sum_{\theta=1}^{L}\frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}d\boldsymbol{y} \\
&\leq L\cdot\max_{\theta\in\Lambda}\frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{Y}^n}[1-\omega(i|\boldsymbol{y})]\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}d\boldsymbol{y} \\
&= L\cdot K_n(\Omega),
\end{aligned} \tag{35}$$

and so, $L_n \leq L$ independently of n. Of course, the above chain of inequalities continues to hold even if the size of Λ varies with n.

2. Discrete–valued sufficient statistics. Suppose that $p_{\theta_i}(\boldsymbol{y}|H_i)$, $i = 0, ..., M-1$, can be represented as

$$p_{\theta_i}(\boldsymbol{y}|H_i) = Q_i(\theta_i, g_i(\boldsymbol{y})),$$

that is, $p_{\theta_i}(\boldsymbol{y}|H_i)$ depends on $\boldsymbol{y}$ only via a sufficient statistic function g_i, which is independent of θ. Suppose further that the supremum of $p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)$ is a maximum, and that the range of g_i is a finite set for every n, i.e., $G_n \stackrel{\triangle}{=} |\{g_i(\boldsymbol{y}) : \boldsymbol{y} \in \mathcal{Y}^n\}| < \infty$. This is the case, for example, with finite–alphabet memoryless sources, where the sufficient statistic g_i is given by the empirical probability distribution and the number of distinct empirical probability distributions G_n is polynomial in n. More generally, finite–alphabet Markov chains also fall in this category. Now, observe that since $g_i(\boldsymbol{y})$ does not take on more than G_n distinct values as $\boldsymbol{y}$ exhausts $\mathcal{Y}^n$ (by assumption), then neither does the maximizer of $p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)$. In other words, the cardinality of the set $\Lambda_n = \{\text{argmax}_\theta p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta) : \boldsymbol{y} \in \mathcal{Y}^n\}$ is at most G_n. Since

$$\max_\theta \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)} \leq \sum_{\theta\in\Lambda_n} \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)},$$

we can repeat the chain of inequalities (35) with the finite summation over θ being taken over Λ_n. Finally, the last equality in eq. (35) is now replaced by an inequality because the maximum over Λ_n never exceeds the supremum over Λ. The conclusion, then, is that in this case $L_n \leq G_n$.

3. Dense grids for smooth parametric families. Example 1 essentially extends to the case of a continuous index set Λ even if the assumptions of Example 2 are relaxed, but then the requirement would be that $p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)$ is sufficiently smooth as a function of θ. Specifically, the idea is to form a sequence of finite grids $\Lambda_n = \{\theta^1, ..., \theta^{l_n}\}$, $\theta^i \in \Lambda$, $i = 1, \ldots, l_n$, that on the one hand, becomes dense in Λ as $n \to \infty$, and on the other hand, its size l_n is subexponential in n. These two requirements can simultaneously be satisfied as long as the classes of sources are not large. Now, if in addition,

$$\gamma_n \stackrel{\triangle}{=} \sup_{\boldsymbol{y}\in\mathcal{Y}^n} \frac{\sup_{\theta\in\Lambda} p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)}{\max_{\theta\in\Lambda_n} p_{\theta_i}(\boldsymbol{y}|H_i)/P_e^*(\theta)} < \infty \tag{36}$$

then again, by a similar chain of inequalities as above, it is easy to show that $L_n \leq l_n\gamma_n$. Thus, the asymptotic minimaxity of $\hat{\Omega}$ can be established if $\{\gamma_n\}$ is subexponential as well, which is the smoothness requirement needed. This requirement on γ_n might be too restrictive, especially if $\mathcal{Y}^n$ is unbounded. Nonetheless, it can sometimes be weakened in such a way that the supremum in (36) is taken merely over a bounded set of very high probability under every possible probability measure of every hypothesis (see, e.g., [14],[5]). This technique of using a grid was also used in [5] in the context of universal decoding. However, in contrast to [5], here the grid is not used in the decision algorithm itself, but only to describe the sufficient condition. Our proposed decision rule continues to be $\hat{\Omega}$, independently of the grid. We will further elaborate on this in Section 4.

Discussion

To gain some more general insight on the conditions under which L_n is subexponential when $\{\Lambda_i\}$ are continuous, observe that the passage from the expression of $K_n(\Omega)$ to that of $\hat{K}_n(\Omega)$ requires that the maximization over Λ and the integration over $\mathcal{Y}^n$ would essentially be interchangeable. To this end, it is sufficient that the integral of

$$\hat{f}_i(\boldsymbol{y}|\theta) \triangleq \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{P_e^*(\theta)}$$

over $\theta \in \Lambda$ would be asymptotically equivalent to $\hat{f}_i(\boldsymbol{y}) = \sup_{\theta\in\Lambda} \hat{f}_i(\boldsymbol{y}|\theta)$ in the exponential scale, uniformly for every $\boldsymbol{y}$ with the possible exception of a set of points whose probability is negligibly small. Since

$$\int_\Lambda \hat{f}_i(\boldsymbol{y}|\theta)d\theta \le \mathrm{Vol}(\Lambda)\cdot \hat{f}_i(\boldsymbol{y}), \tag{37}$$

it is sufficient to require that the converse inequality essentially holds as well (within a subexponetial factor). For the integral of $\hat{f}_i(\boldsymbol{y}|\theta)$ to capture the maximum of $\hat{f}_i(\boldsymbol{y}|\theta)$, there should be a neighborhood of points in Λ, around the maximizer of $\hat{f}_i(\boldsymbol{y}|\theta)$, such that on the one hand, $\hat{f}_i(\boldsymbol{y}|\theta)$ is close to $\hat{f}_i(\boldsymbol{y})$ for every θ in that neighborhood, and on the other hand, the volume of this neighborhood is non-vanishing. In this case, the integral of $\hat{f}_i(\boldsymbol{y}|\theta)$ over Λ is lower bounded by the volume of this neighborhood multiplied by the minimum of $\hat{f}_i(\boldsymbol{y}|\theta)$ within the neighborhood, but this minimum is still fairly close to $\hat{f}_i(\boldsymbol{y})$.

It is interesting to point out that the very same idea serves as the basis of asymptotic methods of Laplace integration techniques [4],[23]. We have deliberately chosen to keep the foregoing discussion informal, but with hopefully clear intuition, rather than giving a formal condition, which might be difficult to verify in general.

It should also be pointed out that some of the techniques used in the above examples were essentially used in [17] to show that the GLRT is asymptotically minimax in the sense of minimizing $\max_{\theta\in\Lambda} P_e(\Omega|\theta)$. This observation indicates that the GLRT is a more pessimistic criterion because performance is not measured relative to the optimum ML decision rule.

3.2 Suboptimal Decision Rules

Although the decision rule $\hat{\Omega}$ is easier to implement and more explicit than the exact minimax decision rule, its implementation is still not trivial. The main difficulty is that it requires an exact closed-form expression of $P_e^*(\theta)$ for every $\theta \in \Lambda$, which is, unfortunately, rarely available.

In some situations, where $P_e^*(\theta)$ decays exponentially and the error exponent function

$$E(\theta) = \lim_{n\to\infty}[-\frac{1}{n}\ln P_e^*(\theta)]$$

is available in closed form, then $\hat{f}_i(\boldsymbol{y})$ can be further approximated by

$$\sup_\theta \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{\exp[-nE(\theta)]}. \tag{38}$$

Clearly, as can be shown using the same techniques as in subsection 3.1. the resulting decision rule inherits the asymptotic minimaxity property of $\hat{\Omega}$ provided that the convergence of $-\frac{1}{n}\ln P_e^*(\theta)$ to $E(\theta)$ is uniform across Λ.

In many other situations, however, even the exact exponential rate function $E(\theta)$ is not available in closed form. Suppose, nonetheless, that there is an explicit expression of an *upper bound* $B(\theta)$ to $P_e^*(\theta)$, which is often the case in many applications. Consider the test statistic

$$\tilde{f}_i(\boldsymbol{y}) = \sup_{\theta\in\Lambda} \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{B(\theta)}, \tag{39}$$

and let $\tilde{\Omega} = \{\tilde{\Omega}_0, \ldots, \tilde{\Omega}_{M-1}\}$, be a decision rule where

$$\tilde{\Omega}_i = \{\boldsymbol{y}:\ \tilde{f}_i(\boldsymbol{y}) \geq \max_{j\neq i} \tilde{f}_i(\boldsymbol{y})\}, \quad i = 0, \ldots, M-1 \tag{40}$$

and where ties are broken arbitrarily. Now, define

$$\tilde{K}_n(\Omega) \stackrel{\triangle}{=} \sup_\theta \frac{P_e(\Omega|\theta)}{B(\theta)} \tag{41}$$

and let $K_n'(\Omega)$ be defined similarly as $\hat{K}_n(\Omega)$ but with the denominator $P_e^*(\theta)$ being replaced by $B(\theta)$, i.e., $\hat{f}_i$ is replaced by $\tilde{f}_i$. Finally, let

$$\tilde{L}_n = \sup_\Omega \frac{K_n'(\Omega)}{\tilde{K}_n(\Omega)} \tag{42}$$

The following theorem gives an upper bound to the error probability associated with $\tilde{\Omega}$.

Theorem 2 *For every $\theta \in \Lambda$,*

$$P_e(\tilde{\Omega}|\theta) \leq \tilde{L}_n K_n B(\theta).$$

Note that $\tilde{L}_n$ can be assessed using the same considerations as discussed in subsection 3.1, and therefore, under certain regularity conditions, it is subexponential similarly to L_n. If, in addition, K_n is subexponential (i.e., if there exists a universal decision rule in the error exponent sense), then Theorem 2 tells us that the exponential decay rate of the error probability associated with $\tilde{\Omega}$ is at least as good as that of the upper bound $B(\theta)$. This opens a variety of possible tradeoffs between guaranteed performance and ease of implementation. Loose bounds typically have simple expressions but then the guaranteed performance might be relatively poor. On the other hand, more sophisticated and tight bounds can improve performance, but then the resulting expression of $B(\theta)$ might be difficult to work with. We shall see a few examples in Section 4.

Proof. First observe that since $P_e(\theta) \leq B(\theta)$ for all θ, then $\tilde{K}_n(\Omega) \leq K_n(\Omega)$ for all Ω. Now, similarly as in the proof of Theorem 1,

$$\tilde{K}_n(\tilde{\Omega}) \leq \tilde{L}_n \inf_{\Omega} \tilde{K}_n(\Omega) \leq \tilde{L}_n \inf_{\Omega} K_n(\Omega) = \tilde{L}_n K_n,$$

and the desired result follows from the definition of $\tilde{K}_n(\Omega)$. □

3.3 Asymptotic Minimaxity Relative to $[P_e^*(\theta)]^\xi$

Returning to the case where $P_e^*(\theta)$ (or at least its asymptotic exponent $E(\theta)$) is available in closed form, it is also interesting to consider the choice $B(\theta) = [P_e^*(\theta)]^\xi$ (or $B(\theta) = e^{-n\xi E(\theta)}$), where $0 \leq \xi \leq 1$. The rationale behind this choice is the following: In certain situations, the competitive minimax criterion w.r.t. $P_e^*(\theta)$ might be too ambitious, i.e., the value of the minimax may grow exponentially with n. Nonetheless, a reasonable compromise of striving to uniformly achieve only a certain fraction ξ of the optimum error exponent, might be achievable. Note that the choice of ξ between zero and unity, gives a spectrum of possibilities that bridges between the GLRT on the one extreme ($\xi = 0$), and the new proposed competitive minimax decision rule on the other extreme ($\xi = 1$). The implementation of the approximate version of this decision rule is not more difficult than that of $\xi = 1$. The only difference is that the denominator of test statistic (38) is replaced by $\exp\{-n\xi E(\theta)\}$ (or, in view of subsection 3.2, Theorem 2, it can even be replaced by $\exp\{-n\xi E_L(\theta)\}$ for some known lower bound $E_L(\theta)$ to $E(\theta)$, if $E(\theta)$ itself is not available in closed form). We propose the following guideline for the choice of ξ: Note that if the competitive minimax value w.r.t. $[P_e^*(\theta)]^\xi$, for a certain value of $\xi > 0$, does not grow exponentially with n, then an error exponent of at least $\xi E(\theta)$ is achieved for all θ. This guarantees that whenever $P_e^*(\theta)$ decays exponentially rapidly (that is, $E(\theta) > 0$), so does the probability of error of the (approximate) minimax decision rule competitive to $[P_e^*(\theta)]^\xi$. We would then like to let ξ be the largest number with this property. More precisely, we wish to select $\xi = \xi^*$, where

$$\xi^* \stackrel{\triangle}{=} \sup\{\xi : \ \limsup_{n\to\infty} \frac{1}{n} \ln K_n^\xi \leq 0\} \tag{43}$$

and

$$K_n^\xi \stackrel{\triangle}{=} \inf_{\Omega} \sup_{\theta\in\Lambda} \frac{P_e(\Omega|\theta)}{[P_e^*(\theta)]^\xi}. \tag{44}$$

In a sense, we can think of the factor ξ^* as the unavoidable cost of uncertainty in θ. Quite clearly, all this is interesting only for cases where $\xi^* > 0$. Fortunately, it turns out that at least in some interesting examples of the composite hypothesis testing problem, it is easy to show that $\xi^* > 0$. One such example, which is analyzed in the Appendix, is the following communication system: Consider the additive Gaussian channel

$$y_t = \theta x_t + z_t, \quad t = 1, 2, \ldots \tag{45}$$

where θ is an unknown gain parameter, and $\{z_t\}_{t\geq 1}$ are i.i.d., zero–mean, Gaussian random variables with variance σ^2. Consider a codebook of two codewords of length n given by $\boldsymbol{x}^0 = (x_1^0, \ldots, x_n^0) = (\sqrt{nP_0}, 0, 0, \ldots, 0)$ and $\boldsymbol{x}^1 = (x_1^1, \ldots, x_n^1) = (0, \sqrt{nP_1}, 0, \ldots, 0)$, where P_0 and P_1 designate the transmission powers associated with the two codewords, which may not be the same. It is demonstrated in the Appendix that while the optimum error exponent of the ML decision rule is given by $E(\theta) = \theta^2(P_0 + P_1)/(8\sigma^2)$, there is a certain decoder, independent of θ, which will be denoted by Ω^0, that achieves an error exponent of $\theta^2 P_0 P_1/[2\sigma^2(P_0 + P_1)]$. Now, for every

$$\xi < \frac{\theta^2 P_0 P_1/[2\sigma^2(P_0 + P_1)]}{\theta^2(P_0 + P_1)/(8\sigma^2)} = \frac{4P_0P_1}{(P_0 + P_1)^2} \tag{46}$$

we have

$$K_n^\xi \leq \sup_{\theta} \frac{P_e(\Omega^0|\theta)}{[P_e^*(\theta)]^\xi} \tag{47}$$

which in turn is of the (non–positive) exponential order of

$$\sup_{\theta} \exp\left\{n\theta^2 \left[\xi - \frac{4P_0P_1}{(P_0 + P_1)^2}\right] \frac{(P_0 + P_1)}{8\sigma^2}\right\} = 1.$$

Therefore, in this case,

$$\xi^* \geq \frac{4P_0P_1}{(P_0+P_1)^2} > 0. \tag{48}$$

The conclusion, therefore, is that the approximate competitive minimax decision rule with $\xi = 4P_0P_1/(P_0+P_1)^2$ is uniformly *at least* as good as Ω^0 for all θ in the error exponent sense. Note that for $P_0 = P_1$, we have $4P_0P_1/(P_0+P_1)^2 = 1$, which implies that $\xi^* = 1$. This means that $E(\theta)$ is universally attainable for orthogonal signals of the same energy. As is shown in the Appendix, in this particular case, even the GLRT attains $E(\theta)$ universally. Another example of theoretical and practical interest, where $\xi^* > 0$ will be discussed in subsection 4.1.

In general, it may not be trivial to compute the exact value of ξ^*. However, it might be possible to obtain upper and lower bounds from lower and upper bounds on K_n^ξ, respectively. Upper bounds on ξ^* would be interesting for establishing fundamental limitations on uniformly achievable error exponents whereas lower bounds yield positive achievability results.

In the foregoing discussion, we demonstrated one way to obtain a lower bound to ξ^* from an upper bound to K_n^ξ. We now conclude this subsection by demonstrating another method, that leads to a single–letter formula of a lower bound to ξ^*, which is tight under the mild regularity conditions described in Examples 1–3 and the Discussion of subsection 3.1. As an example, consider the class of discrete of memoryless sources $\{P_\theta\}$ of a given finite alphabet $\mathcal{Y}$, where θ designates the vector of letter probabilities. Assume further that there are $M = 2$ composite hypotheses, designated by two disjoint subsets, Λ_0 and Λ_1, of this class of sources. In the following derivation, where we make use of the method of types [2], the notation $a_n \doteq b_n$ means that the sequences $\{a_n\}$ and $\{b_n\}$ are of the same exponential order, i.e., $\lim_{n\to\infty} \frac{1}{n} \ln a_n/b_n = 0$. Similarly as in eq. (23), we have

$$\begin{aligned} K_n^\xi &= \frac{1}{2}\sup_\mu \sum_{\mathcal{Y}^n} \min_{i\in\{0,1\}} \int_\Lambda \mu(d\theta)\frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{[P_e^*(\theta)]^\xi} \\ &\leq \sum_{\mathcal{Y}^n} \min_{i\in\{0,1\}} \sup_\theta \frac{p_{\theta_i}(\boldsymbol{y}|H_i)}{[P_e^*(\theta)]^\xi}, \end{aligned} \tag{49}$$

where the inequality is tight in the exponential order under the conditions discussed in subsection 3.1. Now, let $Q_{\boldsymbol{y}}$ denote the empirical PMF (relative frequencies of letters) associated with $\boldsymbol{y}$. Let T_Q denote the type class corresponding to Q, i.e., the set of all n-sequences $\boldsymbol{y} \in \mathcal{Y}^n$ for which $Q_{\boldsymbol{y}} = Q$. Finally, let $\mathcal{Q}^n$ denote the set of all empirical

PMF's of n-sequences over $\mathcal{Y}^n$. Then it is well–known [2] that

$$p_{\theta_i}(\boldsymbol{y}|H_i) = \exp\{-n[H(Q_{\boldsymbol{y}}) + D(Q_{\boldsymbol{y}}||P_{\theta_i})]\} \tag{50}$$

where $H(Q_{\boldsymbol{y}})$ is the empirical entropy of $\boldsymbol{y}$ and $D(Q_{\boldsymbol{y}}||P_{\theta_i})$ is the relative entropy between $Q_{\boldsymbol{y}}$ and P_{θ_i}. Using this and the well-known fact that $|T(Q)| \doteq \exp\{nH(Q)\}$, we now have

$$\begin{aligned}
K_n^\xi &\leq \sum_{Q\in\mathcal{Q}^n} |T(Q)| \cdot \min_{i=0,1} \sup_\theta \frac{\exp\{-n[H(Q) + D(Q||P_{\theta_i})]\}}{\exp\{-n\xi E(\theta)\}} \\
&\doteq \max_Q \min_{i=0,1} \sup_\theta \exp\{n[\xi E(\theta) - D(Q||P_{\theta_i})]\} \\
&= \exp\left\{n[\max_Q \min\{\sup_\theta(\xi E(\theta) - D(Q||P_{\theta_0})), \sup_\theta(\xi E(\theta) - D(Q||P_{\theta_1}))\}]\right\}. \qquad (51)
\end{aligned}$$

For this expression to be subexponential in n, the following condition should be satisfied: For every PMF Q over $\mathcal{Y}$, either $\xi E(\theta) \leq D(Q||P_{\theta_0})$ for all θ or $\xi E(\theta) \leq D(Q||P_{\theta_1})$ for all θ. Equivalently,

$$\xi \leq \min_Q \max\left\{\inf_\theta \frac{D(Q||P_{\theta_0})}{E(\theta)}, \inf_\theta \frac{D(Q||P_{\theta_1})}{E(\theta)}\right\} \tag{52}$$

and so, the r.h.s. is a lower bound to ξ^*. Note, that if Λ_0 and Λ_1 are not *separated away*, and if θ_0 and θ_1 are *unrelated* (in the sense that they may take on values in Λ_0 and Λ_1, respectively, independently of each other), then there exists $Q = Q^*$ for which both numerators of eq. (52) vanish, yet the denominators are strictly positive, and so, $\xi^* = 0$. If, however, θ_0 and θ_1 are related (e.g., θ_1 is some function of θ_0), then ξ^* could be strictly positive as the denominators of eq. (52) may tend to zero with the numerators. A simple example of this is the class of binary memoryless sources (Bernoulli) with $\mathcal{Y} = \{0, 1\}$, where θ designates the probability of '1', $\Lambda_0 = [0, 1/2)$, and $\Lambda_1 = (1/2, 1]$. Again, if θ_0 and θ_1 are unrelated, then $\xi^* = 0$. However, if θ_0 and θ_1 are related by $\theta_0 = 1 - \theta_1$, then $\xi^* = 1$. This is not surprising as the ML decision rule, which achieves $E(\theta)$, is independent of θ in this case.

4 Applications

In this section, we examine the applicability of our approach to two frequently encountered problems of signal processing and communications. We will also compare our method to other commonly used methods, in particular, the GLRT. As mentioned earlier, special attention will be devoted to the problem of universal decoding that arises in coded communication over unknown channels.

4.1 Pattern Recognition Using Training Sequences

Consider the following problem in multiple hypothesis testing, which is commonly studied in statistical methods of pattern recognition, like speech recognition and optical character recognition (see also [31], [8], [15]). There is a model of some parametric family of PDF's $\{q_\phi(\boldsymbol{y}),\ \phi \in \Phi\}$ (e.g., hidden Markov models in the case of speech recognition), and M sources $q_{\phi_i}(\cdot)$, $i = 0, 1, \ldots, M-1$, in this class constitute the M hypotheses to which a given observation sequence $\boldsymbol{z}$ must be classified. For simplicity, let us assume that $M = 2$ and the two sources are a-priori equiprobable. Obviously, if ϕ_i, $i = 0, 1$, were known this would have been a simple hypothesis testing problem. What makes this a composite hypothesis testing problem is that, in practice, ϕ_0 and ϕ_1 are unknown, and instead, we are given two independent training sequences $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, emitted by q_{ϕ_0} and q_{ϕ_1}, respectively. To formalize this in our framework, the entire data set is $\boldsymbol{y} = (\boldsymbol{z}, \boldsymbol{x}_0, \boldsymbol{x}_1)$, the parameter is $\theta = (\phi_0, \phi_1) \in \Phi^2$, and

$$H_0: \quad p_\theta(\boldsymbol{y}|H_0) = q_{\phi_0}(\boldsymbol{z}) q_{\phi_0}(\boldsymbol{x}_0) q_{\phi_1}(\boldsymbol{x}_1),$$

$$H_1: \quad p_\theta(\boldsymbol{y}|H_1) = q_{\phi_1}(\boldsymbol{z}) q_{\phi_0}(\boldsymbol{x}_0) q_{\phi_1}(\boldsymbol{x}_1).$$

In words, under H_i it is assumed that $\boldsymbol{z}$ shares the same parameter as $\boldsymbol{x}_i$, $i = 0, 1$.

Denote by $P_e^*(\phi_0, \phi_1)$ the minimum error probability associated with the simple hypothesis testing problem defined by (ϕ_0, ϕ_1). This is the error attained by LRT, comparing $q_{\phi_0}(\boldsymbol{z})$ to $q_{\phi_1}(\boldsymbol{z})$. Based on the above, our asymptotically competitive minimax decision rule will select the hypothesis H_i for which

$$\hat{f}_i(\boldsymbol{y}) = \max_{\phi_0, \phi_1} \frac{q_{\phi_i}(\boldsymbol{z}) q_{\phi_0}(\boldsymbol{x}_0) q_{\phi_1}(\boldsymbol{x}_1)}{P_e^*(\phi_0, \phi_1)} \qquad i = 0, 1,\ j \neq i$$

is maximum. This is, in general, different from the Bayesian approach [15], where the decision is according to the i that maximizes

$$\max_\phi [q_\phi(\boldsymbol{x}_i) q_\phi(\boldsymbol{z})] \cdot \max_{\phi'} q_{\phi'}(\boldsymbol{x}_j),$$

and from the GLRT [31],[8] used under the Neyman–Pearson criterion, where

$$\max_\phi [q_\phi(\boldsymbol{x}_0) q_\phi(\boldsymbol{z})] / [\max_\phi q_\phi(\boldsymbol{x}_0) \max_\phi q_\phi(\boldsymbol{z})]$$

is compared to a threshold (independently of $\boldsymbol{x}_1$).

As a simple example, consider the case of two Gaussian densities given by

$$q_{\phi_i}(\boldsymbol{z}) = (2\pi)^{-n/2}\exp\{-\frac{1}{2}\sum_{t=1}^{n}(z_t-\phi_i)^2\}, \quad i=0,1$$

where ϕ_0 and ϕ_1 take on values in a certain interval $[-A,A]$, $A>0$, and we are given two training sequences $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ of length m. The exact expression of $P_e^*(\phi_0,\phi_1)$ is given by

$$P_e^*(\phi_0,\phi_1) = Q\left(\frac{\sqrt{n}}{2}|\phi_0-\phi_1|\right),$$

where

$$Q(x) \stackrel{\triangle}{=} \int_x^\infty \frac{du}{\sqrt{2\pi}}e^{-u^2/2}.$$

The asymptotic error exponent associated with $P_e^*(\phi_0,\phi_1)$ is given by

$$E(\phi_0,\phi_1) = \frac{(\phi_0-\phi_1)^2}{8}.$$

Thus, the computation of $\hat{f}_i(\boldsymbol{y})$, with the denominator approximated by $e^{-n(\phi_0-\phi_1)^2/8}$, involves maximization of a quadratic function of ϕ_0 and ϕ_1, which can be carried out in closed form. Specifically, the maximizations associated with $\hat{f}_0(\boldsymbol{y})$ and $\hat{f}_1(\boldsymbol{y})$ are equivalent to the minimizations of

$$\|\boldsymbol{z}-\phi_0\|^2 + \|\boldsymbol{x}_0-\phi_0\|^2 + \|\boldsymbol{x}_1-\phi_1\|^2 - \frac{n}{4}(\phi_0-\phi_1)^2$$

and

$$\|\boldsymbol{z}-\phi_1\|^2 + \|\boldsymbol{x}_0-\phi_0\|^2 + \|\boldsymbol{x}_1-\phi_1\|^2 - \frac{n}{4}(\phi_0-\phi_1)^2,$$

respectively, both over $[-A,A]^2$. At this point, it is important and interesting to distinguish between two cases regarding the relative amount of training data. If $m > n(\sqrt{5}-1)/4$, these two quadratic functions have positive definite Hessian matrices (independently of the data), and hence also have global minima even for $A=\infty$. Therefore, if the absolute values of the true ϕ_0 and ϕ_1 are significantly less than A, then with high probability, these minimizers are also in the interior of $[-A,A]^2$. In this situation, the proposed approximate minimax decision rule, similarly to the GLRT, decides according to whether the sample mean of $\boldsymbol{z}$ is closer to the sample mean of $\boldsymbol{x}_0$ or to the sample mean of $\boldsymbol{x}_1$. If, on the other hand, $m < n(\sqrt{5}-1)/4$, then the Hessian matrix of each one of the above mentioned quadratic forms has a negative eigenvalue, and so, its minimum is attained always at the boundary of $[-A,A]^2$. In this case, the decision rule might be substantially different.

Because of this "threshold effect", and the intuition that attainable error exponents must depend on the amount of training data, this example is an excellent example where it would be advisable to apply an approximate minimax decision rule w.r.t. $[P_e^*(\theta)]^\xi$ for some $\xi < 1$ (cf. subsection 3.3). At the technical side, note that below a certain value of ξ (depending on m/n), each of the quadratic forms to be minimized over (ϕ_0, ϕ_1) (for which now the term $n(\phi_0 - \phi_1)^2/4$ is multiplied by ξ) is again guaranteed to have a positive definite Hessian matrix. As for a lower bound to ξ^* for his problem, it is not difficult to show, using the Chernoff bound, that the GLRT (for $A = \infty$) attains an error exponent of $(\phi_0 - \phi_1)^2/[8(1 + n/(2m)]$. It then follows that, in this case, $\xi^* \geq 1/[1 + n/(2m)]$.

4.2 Universal Decoding

The problem of universal decoding is frequently encountered in coded communication: When the channel is unknown, the ML decoder cannot be implemented and a good decoder is sought that does not depend on the unknown values of the channel parameters. We first provide a brief description of the problem and prior work on the subject, and then examine our approach in this context.

Consider a family of vector channels $\{W_\theta(\boldsymbol{y}|\boldsymbol{x}),\ \theta \in \Lambda\}$, where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is the channel input, $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ is the observed channel output, and θ is the index (or the parameter) of the channel in the class. A block code $\mathcal{C} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\} \subset \mathcal{X}^n$ of length n and rate R is a collection of $M = 2^{nR}$ vectors of length n, which represent the set of messages to be transmitted across the channel. Upon transmitting one of the M messages $\boldsymbol{x}^i$, a vector $\boldsymbol{y}$ is received at the channel output, under the conditional PDF $W_\theta(\cdot|\boldsymbol{x}^i)$. The decoder, which observes $\boldsymbol{y}$ and knows $\mathcal{C}$, but does not know θ, has to decide which message was transmitted. This is, of course, a composite hypothesis problem with multiple hypotheses, where the same parameter value θ corresponds to all hypotheses and $p_\theta(\boldsymbol{y}|H_i) = W_\theta(\boldsymbol{y}|\boldsymbol{x}^i)$, $i = 0, 1, \ldots, M-1$.

It is well-known [2] that for discrete memoryless channels (DMC's) (see also [14] for Gaussian memoryless channels) and more generally, for finite-state (FS) channels [30], [11], there exist universal decoders in the *random coding* sense. Specifically, the exponential decay rate of the average error probability of these universal decoders, w.r.t. the ensemble of randomly chosen codes, is the same as that of the average error probability obtained by the optimum ML decoder. Universality in the random-coding sense does not imply that for a specific code the decoder attains the same performance as the optimal ML decoder, nor does it imply that there exists a specific code for which the universal decoder has good performance.

In a recent work, [5], these results have been extended in several directions. First, the universality in the random coding sense has been generalized in [5] to arbitrary indexed classes of channels obeying some mild regularity conditions on smoothness and richness. Secondly, under somewhat stronger conditions, referred to as *strong separability* in [5], the convergence rate towards the optimal random coding exponent is uniform across the index set [5, Theorem 2], namely,

$$\lim_{n\to\infty} \sup_{\theta} \frac{1}{n} \log \frac{\bar{P}_e(\Omega|\theta)}{\bar{P}_e^*(\theta)} = 0,$$

where $\bar{P}_e(\Omega|\theta)$ is the random-coding average error probability associated with the universal decoder Ω, and $\bar{P}_e^*(\theta)$ is the one associated with the optimum ML decoder for W_θ. Finally, it was shown that, under the same condition, there exists a sequence of specific codes $\{\mathcal{C}_n,\ n = 1, 2, ...\}$, for which the universal decoder of [5] achieves the random coding error exponent of the ML decoder uniformly in θ.

The existence of a universal decoder in the error exponent sense uniformly in θ, for both random codes and deterministic codes, obviously implies that both

$$K_n = \inf_{\Omega} \sup_{\theta} \frac{\bar{P}_e(\Omega|\theta)}{\bar{P}_e^*(\theta)} \quad \text{and} \quad K_n[\mathcal{C}_n] = \inf_{\Omega} \sup_{\theta} \frac{P_e(\Omega|\theta, \mathcal{C}_n)}{\bar{P}_e^*(\theta)}, \tag{53}$$

where $P_e(\Omega|\theta, \mathcal{C}_n)$ is the probability of error for a specific code $\mathcal{C}_n$ (in the same sequence of codes as in [5]) are subexponential in n (that is, $\xi^* = 1$ for both). Therefore, similarly as in the derivation in Section 2, it is easy to show that the following decision rule is universal (relative to the random coding exponent) in both the random coding sense and in the deterministic coding sense: Decode the message $\boldsymbol{x}^i$ as the one that maximizes the quantity

$$\hat{f}(\boldsymbol{x}^i, \boldsymbol{y}) = \max_{\theta} \frac{W_\theta(\boldsymbol{y}|\boldsymbol{x}^i)}{\bar{P}_e^*(\theta)}. \tag{54}$$

This decoder can be further simplified by its asymptotically equivalent version

$$\log \hat{f}(\boldsymbol{x}^i, \boldsymbol{y}) \approx \max_{\theta}[\log W_\theta(\boldsymbol{y}|\boldsymbol{x}^i) + nE_r(R, \theta)], \tag{55}$$

where $E_r(R,\theta) = \lim_{n\to\infty} -\frac{1}{n} \log \bar{P}_e^*(\theta)$ (whenever the limit exists) is the asymptotic exponent of the average error probability (random-coding exponent [6]) associated with W_θ.

The latter version is typically more tractable since, as explained earlier, explicit closed-form expressions are available much more often for the random coding error exponent function than for the average error probability itself. For example, in the case of DMC's, Gallager's reliability function provides the exact behavior (and not only a lower bound) to the random-coding error exponent [7]. It should be kept in mind, however, that in the case of DMC's, the maximum mutual information (MMI) universal decoder [2], which coincides with the GLRT decoder for fixed composition codes, also attains $E_r(R,\theta)$ for all θ, both in the random coding sense and in the deterministic coding sense. Nevertheless, this may no longer be true for more general families of channels. It should be stressed, however, that the existence of a universal decoder (55) in the deterministic coding sense w.r.t. the random coding error exponent $E_r(R,\theta)$ does not imply that there exists one with the same property w.r.t. the ML–decoding error exponent of the same sequence deterministic codes, that is, $P_e^*(\theta|\mathcal{C}_n)$.

It is important to emphasize also that the universal decoder of eq. (55) is much more explicit than the one proposed in [5] as it avoids the need of employing many decoding lists in parallel, each one corresponding to one point in a dense grid (whose size grows with n) in the index set, as proposed in [5].

As an additional benefit of this result, more understanding can be gained regarding the performance of GLRT, which is so commonly used when the channel is unknown. We have already mentioned that in some cases (e.g., the class of DMC's) the GLRT performs equally well as the universal decoder proposed herein. In some other cases, this is trivially so, simply because the two decoders coincide. For example, if $\bar{P}_e^*(\theta)$ happens to be independent of θ in a certain instance of the problem, then the GLRT is universal, simply because it coincides with (54). For example, consider an additive channel with a jammer signal [14] parameterized by θ, i.e., $y_t = x_t + z_t + j_t(\theta)$, where z_t is additive noise (with known statistics) and $j_t(\theta)$ is a deterministic jammer signal characterized by θ (e.g., a sine–wave with a certain amplitude, frequency, and phase). Here, when θ is known, $j_t(\theta)$ can be subtracted from y_t and so, $\bar{P}_e^*(\theta)$ is the same as for the channel $y_t = x_t + z_t$, which in turn is independent of θ. Another example, in a continuous time setting, is associated with a constant energy, orthogonal signal set given by sine waves at different frequencies (FSK), transmitted via an additive white Gaussian channel (AWGN) with an unknown all–pass filter parameterized by θ. Since the signals remain essentially orthogonal, and with the same energy, even after passing the all–pass filter, $P_e^*(\theta)$ is the probability of error of an orthogonal system in the

AWGN, essentially independently of θ (assuming sufficiently long signaling time).

Perhaps one of the most important models where the universal decoder (54) should be examined is the well–known model of the Gaussian intersymbol-interference (ISI) channel defined by

$$y_t = \sum_{i=0}^{k} h_i x_{t-i} + z_t, \tag{56}$$

where $(h_0, ..., h_k) = \theta$ is the vector of unknown ISI coefficients and $\{z_t\}$ is zero–mean Gaussian white noise with variance σ^2 (known or unknown). The problem of channel decoding with unknown ISI coefficients has been extensively investigated, and there are many approaches to its solution, most of which are on the basis of the GLRT. As mentioned earlier, the results of [5] imply that universal decoding, in the random coding sense, is possible for this class of channels. Therefore the competitive–minimax decoder, proposed herein, as well as its asymptotic approximation,[5] is universal as well in the random coding error exponent sense.

In addition to the random-coding universality, it is especially appealing, in the case of the ISI channel, to examine the performance of our decoder when it is directed to asymptotic minimaxity w.r.t. a specific code. In other words, we wish to implement the same decoder as in eq. (54), but with the denominator being replaced by the probability of error associated with a specific code.

To demonstrate the decoding algorithm explicitly in this case, let us consider, for the sake of simplicity, a codebook $\mathcal{C}$ of two codewords, $\boldsymbol{x}^0$ and $\boldsymbol{x}^1$, and let $\boldsymbol{u} = (u_1, \ldots, u_n) \stackrel{\triangle}{=} \boldsymbol{x}^0 - \boldsymbol{x}^1$. If the ISI channel were known, then the probability of error associated with optimum ML decoding would have been of the exponential order of $\exp\{-\sum_t [\sum_j h_j u_{t-j}]^2/(8\sigma^2)\}$, where the numerator in the exponent is the Euclidean distance between the two codewords after having passed the ISI filter (neglecting some edge effects at the beginning of the block). Let us suppose also that one knows a-priori that the ISI filter is of limited energy, i.e., $\sum_{i=0}^k h_i^2 \leq S$, where $S > 0$ is given. Then, our approximate competitive–minimax decoder (for $\xi = 1$), in this case, picks the codeword $\boldsymbol{x}^i = (x_1^i, \ldots, x_n^i)$, $i = 0, 1$, that minimizes the expression

$$\min_{h_0,\ldots,h_k} \left\{ \frac{1}{2\sigma^2} \sum_t (y_t - \sum_{j=0}^{k} h_j x_{t-j}^i)^2 - \frac{1}{8\sigma^2} \sum_t [\sum_{j=0}^{k} h_j u_{t-j}]^2 \right\} \tag{57}$$

subject to the constraint $\sum_{i=0}^k h_i^2 \leq S$. This is a standard quadratic minimization problem

[5]cf. Example 3 and Discussion in subsection 3.1.

and the minimizing vector $\theta = (h_0, \ldots, h_k)$ of ISI filter coefficients is given by solving the following set of linear equations

$$(G + \lambda I)\theta = g, \tag{58}$$

where I is the $(k+1) \times (k+1)$ identity matrix, λ is a Lagrange multiplier chosen so as to satisfy the energy constraint,

$$g = (\frac{1}{n}\sum_t x_t y_t, \frac{1}{n}\sum_t x_{t-1} y_t, \ldots, \frac{1}{n}\sum_t x_{t-k} y_t)^T,$$

and G is a $(k+1) \times (k+1)$ matrix whose (i,j)th entry is given by

$$\frac{1}{n}\sum_t x_{t-i} x_{t-j} - \frac{1}{4n}\sum_t u_{t-i} u_{t-j}.$$

For low–rate codebooks of size larger than 2, a similar idea can still be used with $P_e^*(\theta)$ being approximated using the union bound, which is given by the pairwise error probability as above, multiplied by the codebook size M. However, this should be done with some care as the pair of codewords $(\boldsymbol{x}^i, \boldsymbol{x}^j)$ that achieves the minimum distance, $\sum_t [\sum_l h_l (x_{t-l}^i - x_{t-l}^j)]^2$, may depend on the filter coefficients. For higher rates, where the union bound is not tight in the exponential scale, more sophisticated bounds must be used.

It is interesting to note that the existence of a universal decoder in the error exponent sense for a specific orthogonal code, can be established using eq. (27). For example, consider the channel defined by $y_t = \theta x_t + z_t$, where $\{z_t\}$ are zero–mean, i.i.d. Gaussian random variables, and θ is an unknown constant. Suppose that $\Lambda = \{a, -a\}$ for some known constant $a > 0$, and the codebook consists of two orthogonal codewords, $\boldsymbol{x}_0 = (x_1^0, \ldots, x_n^0)$ and $\boldsymbol{x}_1 = (x_1^1, \ldots, x_n^1)$. It can easily be seen that for every $\boldsymbol{y} = (y_1, \ldots, y_n)$, the minimax and maximin values at the numerator and denominator of (27) are the same. Thus, $K_n \leq |\Lambda| = 2$ and the existence of a universal decoder is established. This example can be extended to the case of a larger orthogonal code, and for any symmetric set Λ. Also, it can be observed that in this case, the GLRT is a universal decoder. Interestingly, when the codewords are not orthogonal the minimax and maximin values are not equal, and this technique cannot be used to determine whether or not a universal decoder exists. In this case, as is shown in the Appendix, there is a uniformly better decoder than the GLRT [14]. Unfortunately, even that decoder is not universal in the error exponent sense for every specific code.

5 Conclusion and Future Research

In this paper, we proposed and investigated a novel minimax approach to composite hypothesis testing with applications to problems of classification and to universal decoding. The main idea behind this approach is to minimize (or, to approximate the minimizer of) the worst–case loss in performance (in terms of error probability) relative to the optimum ML test that assumes knowledge of the parameter values associated with the different hypotheses. The main important property of the proposed decision rule is that, under certain conditions, it is universal in the error exponent sense whenever such a universal decision rule at all exists. When it is not universal in the error exponent sense, it means that such a universal decision rule does not exist. We studied the properties of the proposed competitive–minimax decision rule, first in the general level, and then in some more specific examples. One of the interesting properties of the proposed decision rule is that, in general, it might be randomized and this is different from the classical solutions to the hypothesis testing problem.

Future research will focus on further studying the properties of our proposed decision rule, mostly in applications of practical interest. Specifically, in the context of universal decoding, more understanding is left to be desired regarding considerations of code design for universal decoding. Tradeoffs between performance and ease of implementation, as discussed in the paper, will also receive more attention in the future. Another direction to be investigated is in defining a minimax criterion that is suitable to the Neyman–Pearson setting of the composite hypothesis testing problem, and deriving a corresponding decision rule. This could be useful in problems of signal detection and model order estimation.

Appendix

In this appendix, we demonstrate the suboptimality of the GLRT in a very simple example. Consider the additive Gaussian channel

$$y_t = \theta x_t + z_t, \quad t = 1, 2, \ldots \tag{A.1}$$

where θ is an unknown gain parameter, and $\{z_t\}_{t\geq 1}$ are i.i.d., zero–mean, Gaussian random variables with variance σ^2. Suppose that our codebook consists of two codewords of length n given by $\boldsymbol{x}^0 = (x_1^0, \ldots, x_n^0) = (\sqrt{nP_0}, 0, 0, \ldots, 0)$ and $\boldsymbol{x}^1 = (x_1^1, \ldots, x_n^1) =$

$(0, \sqrt{nP_1}, 0, \ldots, 0)$, where P_0 and P_1 designate the transmission powers associated with the two codewords, which may not be the same.[6] Now, the GLRT picks the codeword $\boldsymbol{x}^i$, $i = 0, 1$, that minimizes $\min_\theta \sum_{t=1}^n (y_t - \theta x_t^i)^2$, which is equivalent to deciding according to $\min_\theta \sum_{t=1}^2 (y_t - \theta x_t^i)^2$, since all coordinates of both codewords vanish for $t \geq 3$. Thus, the problem is actually in two dimensions. Referring to Fig. 1, the GLRT projects the vector (y_1, y_2) onto the directions of the two–dimensional vectors formed by the first two coordinates of $\boldsymbol{x}^0$ and $\boldsymbol{x}^1$ (namely, $(1, 0)$ and $(0, 1)$, respectively), and decides according to the smaller between the distances from (y_1, y_2) to the vertical axis and to the horizontal axis of the coordinate system. In other words, the GLRT decides in favor of $\boldsymbol{x}^0$ or $\boldsymbol{x}^1$ according to whether $|y_1| \leq |y_2|$ or $|y_1| > |y_2|$. Thus, the boundaries between the two decision regions are straight lines through the origin at slopes of ± 45 degrees. Accordingly, the distances from $\theta \cdot (x_1^0, x_2^0)$ and $\theta \cdot (x_1^1, x_2^1)$ to these lines dictate the error probability (refer to the dashed lines in Fig. 1). Specifically, the distance from $(\theta\sqrt{nP_0}, 0)$ to each of the 45 degree boundary lines is $\theta\sqrt{nP_0/2}$ and the distance from $(0, \theta\sqrt{nP_1})$ to the same lines is $\theta\sqrt{nP_1/2}$. It is easy to see then (by rotating the coordinate system by 45 degrees) that the error event given $\boldsymbol{x}^i$ is equivalent to the event that either $U > \theta\sqrt{nP_i/2}$ or $V > \theta\sqrt{nP_i/2}$ (exclusively) $i = 0, 1$, where U and V are independent, zero–mean Gaussian random variables, each with variance σ^2. The probability of error is then given by

$$\begin{aligned} P_e^{GLRT}(\theta) &= \frac{1}{2}\left[2Q\left(\theta\sqrt{\frac{nP_0}{2\sigma^2}}\right) - 2Q^2\left(\theta\sqrt{\frac{nP_0}{2\sigma^2}}\right)\right] + \\ &\quad + \frac{1}{2}\left[2Q\left(\theta\sqrt{\frac{nP_1}{2\sigma^2}}\right) - 2Q^2\left(\theta\sqrt{\frac{nP_1}{2\sigma^2}}\right)\right] \qquad \text{(A.2)} \end{aligned}$$

which is of the exponential order of $\exp\{-n\theta^2 \min\{P_0, P_1\}/(4\sigma^2)\}$.

It is interesting to observe that one can do better than the GLRT, when θ is unknown, by using a decoder that selects the message i for which $\min_\theta \sum_{t=1}^2 (x_t^i - \theta y_t)^2$ is smaller (see [14]), namely, by projecting the vector formed by the first two coordinates of each $\boldsymbol{x}^i$ in the direction of the first two coordinates of $\boldsymbol{y}$. In this case, the boundary between the two decision regions is a pair of straight lines through the origin whose distances to (x_1^0, x_2^0) and to (x_1^1, x_2^1) are the same (the slopes of these lines are $\pm\sqrt{P_1/P_0}$). Elementary geometrical considerations, similar to the above (and the union bound) lead to the result that the

[6] Clearly, every orthogonal code of two codewords can be transformed, by an appropriate orthonormal transformation, to this form. If the original code is not orthogonal, the first coordinate of $\boldsymbol{x}^1$ might be nonzero as well, yet the extension of this example of the suboptimality of the GLRT is still valid.

error probability, in this case, is of the exponential order of $\exp\{-n\theta^2 P_0 P_1/[2\sigma^2(P_0+P_1)]\}$, which is strictly better than that of the GLRT for every non–zero value of θ and for every orthogonal code of two codewords, provided that $P_0 \neq P_1$.

Finally, to complete the picture, consider the ML decision rule. Since the Euclidean distance between $\theta\boldsymbol{x}^0$ and $\theta\boldsymbol{x}^1$ is $\theta\sqrt{n(P_0+P_1)}$, the error probability of ML decoding is of the exponential order of $\exp\{-n\theta^2(P_0+P_1)/(8\sigma^2)\}$, which is strictly better than both previously mentioned exponents, again, provided that $P_0 \neq P_1$.

Note that, in a random coding regime, where P_0 and P_1 are random variables, these exponential error bounds should be averaged w.r.t. the joint ensemble of P_0 and P_1, and so, the random coding error exponent of the GLRT might be strictly inferior to that of the latter universal decoding rule.

Acknowledgement

We are grateful to the anonymous reviewers for their helpful comments.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. AC–19, no. 6, pp. 716–723, December 1974.

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

[3] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT–29, no. 6, pp. 783–795, November 1973.

[4] N. G. de Bruijn, "Asymptotic methods in analysis," Dover Publications, Inc., New York, 1981.

[5] M. Feder and A. Lapidoth, "Universal decoders for channels with memory," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1726–1745, September 1998.

[6] R. G. Gallager, *Information Theory and Reliable Communications*, John Wiley & Sons, 1968.

[7] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, pp. 244-246, March 1973.

[8] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically–observed statistics," *IEEE Trans. Inform. Theory*, vol. IT–35, no. 2, pp. 401–408, March 1989.

[9] C. W. Helstrom, *Statistical Theory of Signal Detection*, Oxford, Pergamon Press, 1968.

[10] W. Hoeffding, "Asymptotically optimal test for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, 1965.

[11] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1746–1755, September 1998.

[12] E. L. Lehmann, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959.

[13] E. Levitan and N. Merhav, "A competitive Neyman–Pearson approach to universal hypothesis testing with applications," submitted to *IEEE Trans. Inform. Theory*, December 2000.

[14] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. IT-39, no. 4, pp. 1261–1269, July 1993.

[15] 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.

[16] 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. IT-35, no. 5, pp. 1014–1019, September 1989.

[17] 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.

[18] N. Merhav and J. Ziv, "A Bayesian approach for classification of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-37, no. 4, pp. 1067–1071, July 1991.

[19] A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill Series in Electrical Engineering, third edition, 1991.

[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[21] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[22] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, 1970.

[23] G. Schwarz, "Estimating the dimension of a model", *The Annals of Stat.*, vol. 6, no. 20, pp. 461–464, 1978.

[24] M. Sion, "On general minimax theorems," *Pac. J. Math.*, vol. 8, pp. 171–176, 1958.

[25] C. C. Tappert, Y. Suen, and T. Wakahara, "The state-of-the-art in on-line hand-written recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 787-808, August 1990.

[26] G. Tusnády, "On asymptotically optimal tests," *Ann. Statist.*,vol. 5, no. 2, pp. 385–393, 1977.

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

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

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

[30] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.

[31] J. Ziv, "On classification with empirically–observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. IT–34, no. 2, pp. 278–286, March 1988.

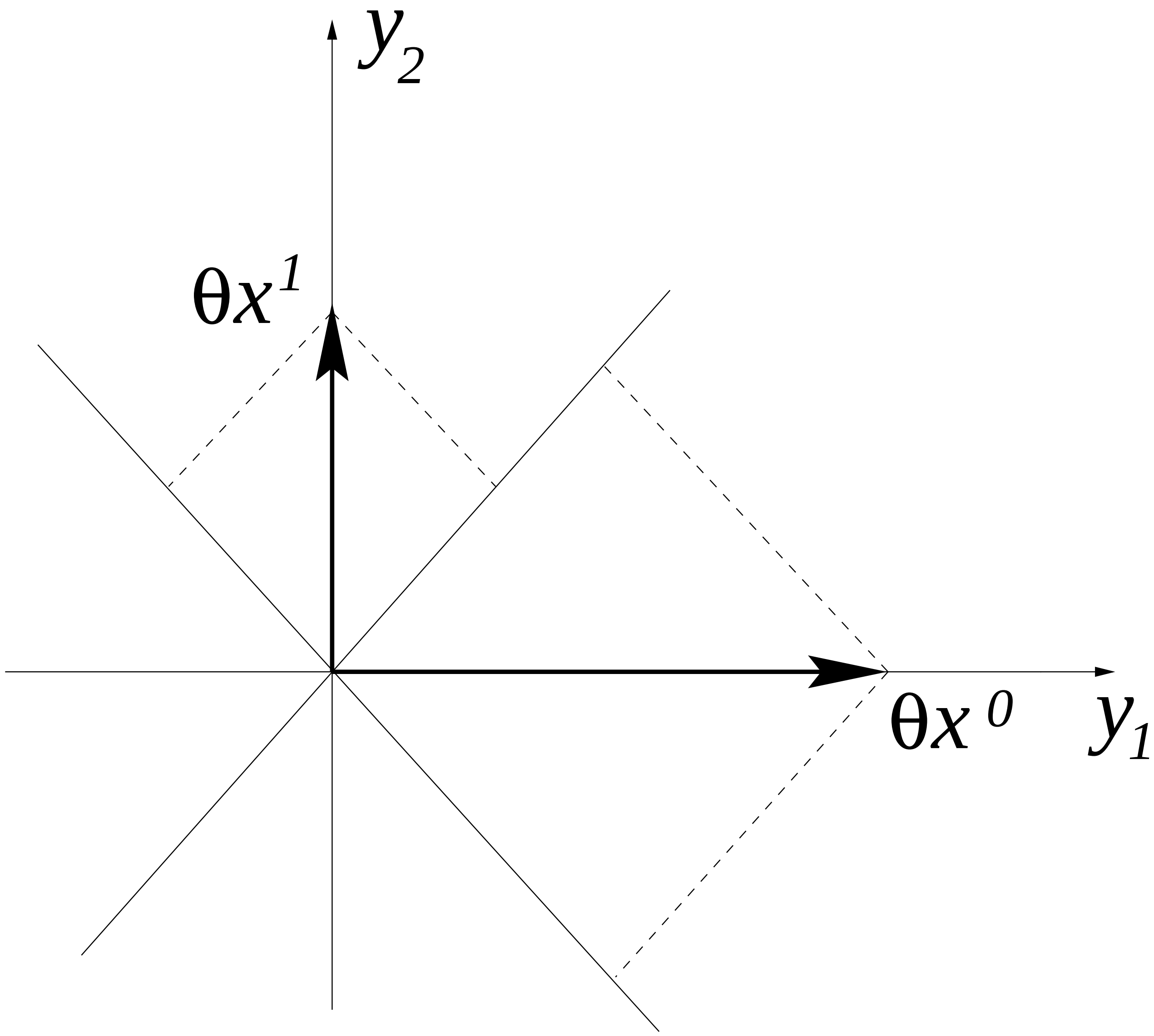

Figure 1: Geometric illustration of the GLRT for two orthogonal codewords.