[9] R. Ahlswede and Z. Zhang, "An identity in combinatorial extremal theory," *Adv. in Math.*, vol. 80, no. 2, pp. 137–151, Apr. 1990.
[10] R. Ahlswede and Z. Zhang, "On cloud-antichains and related configurations," *Discrete Math.*, vol. 85, pp. 225–245, 1990.
[11] A. E. Brouwer, J. B. Shearer, N. J. A. Sloane, and W. D. Smith, "A new table of constant weight codes," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1334–1380, Nov. 1990.

When is the Generalized Likelihood Ratio Test Optimal?

Ofer Zeitouni, *Senior Member, IEEE*, Jacob Ziv, *Fellow, IEEE*, and Neri Merhav, *Member, IEEE*

***Abstract*—The generalized likelihood ratio test (GLRT), which is commonly used in composite hypothesis testing problems, is investigated. Conditions for asymptotic optimality of the GLRT in the Neyman–Pearson sense are studied and discussed. First, a general necessary and sufficient condition is established, and then based on this, a sufficient condition, which is easier to verify, is derived. A counterexample, where the GLRT is not optimal, is provided as well. A conjecture is stated concerning the optimality of the GLRT for the class of finite-state sources.**

***Index Terms*—Hypothesis testing, generalized-likelihood ratio test, maximum-likelihood test, error exponent, Neyman–Pearson criterion, large deviations.**

I. Introduction

Consider the following prototype classification problem. Let $x^n = (x_1, \cdots, x_n)$ be a sequence of observations which take values in a finite set $\boldsymbol{X}$ with cardinality X. It is assumed that x^n has been drawn from a probabilistic source P which is either P_0 or P_1, i.e., P_0 and P_1 are probability measures on the space $\Omega = \boldsymbol{X}^\infty$ of all infinite sequences and we identify the marginal of P_0 under the coordinate map $\omega \to x^n$. The classification problem is that of deciding, upon observing x^n, whether the true underlying source is P_0 or P_1. Throughout the sequel, $P_i(x^n)$, $i = 0, 1$, will denote the probability of a string x^n under P_i. Similarly, $P_i(F)$ will denote the probability of an event F under either source. A decision rule is a subset Λ_n of the sample space $\boldsymbol{X}^n$ such that if $x^n \in \Lambda_n$, then x^n is classified as being drawn from $P = P_1$, otherwise it is classified as $P = P_0$. The probability of error of the first kind (false alarm), associated with Λ_n, is defined as $P_0(\Lambda_n)$, i.e., the probability of deciding $P = P_1$ while P_0 is the true source. Similarly, the probability of error of the second kind (misdetection) is defined as $P_1(\bar{\Lambda}_n)$, where $\bar{\Lambda}_n$ is the set complementary to Λ_n.

For known sources P_0 and P_1, the classical Neyman–Pearson approach [1] suggests the following optimality criterion: Among all decision rules Λ_n yielding

$$P_0(\Lambda_n) \le 2^{-\lambda n}, \tag{1.1}$$

for a given $\lambda > 0$, find the one which minimizes $P_1(\bar{\Lambda}_n)$. It is well known that the solution to this problem is given by the likelihood ratio test (LRT), i.e.,

$$\Lambda_n^{\mathrm{LRT}} = \left\{ x^n : \frac{1}{n} \log P_1(x^n) - \frac{1}{n} \log P_0(x^n) \ge T_n(\lambda) \right\}, \tag{1.2}$$

where logarithms throughout the correspondence will be taken to the base 2 and the threshold function $T_n(\lambda)$, defined such that $P_0(\Lambda_n^{\mathrm{LRT}}) = 2^{-\lambda n}$, depends, in general, also on n, λ, P_0, and P_1.

Suppose next that P_0 is still known but P_1 is unknown except that it is a member of some subclass $\boldsymbol{P}$ of the stationary and ergodic sources, namely, a simple hypothesis $P = P_0$ is tested against a composite alternative $P \in \boldsymbol{P}$. Clearly, in this case, the LRT (1.2) is no longer applicable. It is common to use, in this situation, the generalized-likelihood ratio test (GLRT) [2, p. 1, pp. 86–96], given by

$$\Lambda_n^{\mathrm{GLRT}} = \left\{ x^n : \frac{1}{n} \log \sup_{P_1 \in \boldsymbol{P}} P_1(x^n) - \frac{1}{n} \log P_0(x^n) \ge T_n(\lambda) \right\}. \tag{1.3}$$

In [3]–[6], the GLRT with $T_n(\lambda) = \lambda$ was investigated in several concrete composite hypotheses testing problems, where $\boldsymbol{P}$ was the subclass of all Markov sources up to a given order (including zero order, i.e., memoryless sources). In [7], more generally, a parametric class $\boldsymbol{P}$ from an exponential family was assumed. In fact, all these studies have shown that the GLRT is asymptotically optimal under a modified version of the Neyman–Pearson criterion, defined as follows.

M: Among all sequences of decision rules $\{\Lambda_n\}_{n \ge 1}$ that do not depend on the unknown P_1 and at the same time satisfy

$$\limsup_{n \to \infty} \frac{1}{n} \log P_0(\Lambda_n) < -\lambda \tag{1.4}$$

select a sequence that maximizes the second kind error exponent, $-\limsup_{n \to \infty} n^{-1} \log P_1(\bar{\Lambda}_n)$, uniformly for all $P_1 \in \boldsymbol{P}$.

In other words, the GLRT (1.3), which is independent of P_1, performs in [3]–[7] asymptotically as well as the optimal LRT (1.2), which in turn depends upon P_1. However, in [8], where the class $\boldsymbol{P}$ of alternatives was extended to the set of all finite-state (FS) sources with no more than S states, it was not the GLRT that was proved asymptotically optimal, but a generalized version of a test proposed by Hoeffding (see [9] and Lemma 1 below), where practical implementation turned out to be simplified by the use of the Lempel–Ziv (LZ) algorithm [10]. However, it is not clear from the results of [8] whether the GLRT is still optimal in this case.

In light of these results, the goal of this correspondence is to study relations between the GLRT and the optimal test under the criterion *M* previously defined. In particular, we are interested in establishing conditions on the class $\boldsymbol{P}$ under which the GLRT is asymptotically optimal in that sense. We first derive a necessary and sufficient condition for asymptotic optimality of the GLRT. Since this condition is, in general, difficult to check, we further derive a sufficient condition, which is easier to verify, and demonstrate that it holds in many interesting special cases, including [3]–[7] as well as other cases. On the other hand, we provide a counterexample to demonstrate that the GLRT is *not* always asymptotically optimal. The question whether the GLRT is asymptotically optimal in the case that $\boldsymbol{P}$ is the set of all FS

Manuscript received September 18, 1990. This work was presented at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991. This work was performed in part while the authors were visiting AT & T Bell Laboratories, Murray Hill, NJ.

The authors are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9200885.

sources [8], remains open, although we conjecture that the answer is positive, cf., the remark at the end of Section IV.

As a final remark, we point out that although merely finite-alphabet processes are considered here, the results are readily extended to more general alphabets if one replaces combinatorial bounds by related large deviations bounds and modifies the optimality criterion appropriately. The interested reader is referred to [11] for a precise description of this extension. Also, by using this approach, one may also use the empirical process (provided that $\boldsymbol{P}$ is a set with sufficiently "nice" ergodic properties) and thus, bypass the Markov approximation approach that is used here (see (2.1) and (2.2)).

II. Preliminaries

We shall assume that every source in $\boldsymbol{P}$ can be approximated, to an arbitrary degree of accuracy, by a Markov source of a sufficiently high-order l. Specifically, for every source $P \in \boldsymbol{P}$ and an infinite string $x = (x_1, x_2, \cdots)$ define

$$\rho_l(x) = \sup_{n>l} \left|\log P(x_n|x^{n-1}) - \log P(x_n|x_{n-l}^{n-1})\right|, \qquad l = 1, 2, \cdots, \quad (2.1)$$

where x_i^j denotes the segment $(x_i, x_{i+1}, \cdots, x_j)$ for $j > i$, $P(x_n|x^{n-1})$ is the conditional probability of x_n given the entire past, and $P(x_n|x_{n-l}^{n-1})$ is the conditional probability of x_n given the l preceding letters. It will be assumed that for every $\varepsilon > 0$ there exists a sufficiently large l such that $\rho_l(x) \le \varepsilon$ almost surely. This is equivalent to the condition that for every string x^n with nonzero probability,

$$2^{-\varepsilon n} \prod_{i=1}^{n} P(x_i|x_{i-l}^{i-1}) \le P(x^n) \le 2^{\varepsilon n} \prod_{i=1}^{n} P(x_i|x_{i-l}^{i-1}). \quad (2.2)$$

Since ε is assumed independent of x^n, (2.2) tells us we can approximate P by an lth-order Markov process at the expense of degrading the exponential of rate of the second kind error probability, namely $-\limsup_{n\to\infty} \log P_1(\overline{\Lambda}_n)$, by no more than ε. Therefore, we shall throughout consider sources in $\boldsymbol{P}$ as stationary, ergodic, lth-order Markov sources, keeping in mind that the discussion next will focus, in the general case, on ε-optimal rather than optimal error exponents.

Remark: Rather than using lth-order Markov sources for approximating sources in $\boldsymbol{P}$ one may use, more generally, unifilar FS sources having sufficiently many states (see Appendix) with a straightforward extension of (2.1), (2.2), and the forthcoming derivation. This is advantageous if one wishes to approximate general FS sources [8] as demonstrated in the Appendix. Note, however, that although the GLRT is asymptotically optimal, as mentioned earlier, when $\boldsymbol{P}$ is the class of *all* lth order Markov measures or the class of all unifilar FS sources with X^l states, it is not necessarily so when $\boldsymbol{P}$ is the *subset* of sources, in this class, which approximate (in the sense of (2.2)) all general FS sources with a given number of states.

Let $s_i = x_{i-l}^{i-1}$ denote the state at time instant i. For simplicity, we shall assume that the initial state $s_1 = x_{1-l}^{0}$ is fixed. Define the empirical joint probability of a letter $u \in \boldsymbol{X}$ and a state $\nu \in \boldsymbol{X}^l$ as

$$q_x^n(u, \nu) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, u, s_i, \nu), \quad (2.3)$$

where $\delta(x_i, u, s_i, \nu)$ is the indicator function for $x_i = u$ jointly with $s_i = \nu$. The empirical distribution Q_x^n is defined as the matrix $\{q_x^n(u, \nu)\}_{u\in\boldsymbol{X}, \nu\in\boldsymbol{X}^l}$. The type T_x^n of x^n is the set of all strings $y^n \in \boldsymbol{X}^n$ with $Q_y^n = Q_x^n$. The empirical conditional probability of a letter u given a state ν is defined as

$$q_x^n(u|\nu) = \frac{q_x^n(u, \nu)}{\sum_{u\in\boldsymbol{X}} q_x^n(u, \nu)}, \quad (2.4)$$

with the convention that if the denominator (and hence, also the numerator) is zero, then $q_x^n(u|\nu)$ is set to zero as well. Next, define the empirical conditional entropy associated with x^n as

$$H(Q_x^n) = -\sum_{u\in\boldsymbol{X}} \sum_{\nu\in\boldsymbol{X}^l} q_x^n(u, \nu) \log q_x^n(u|\nu), \quad (2.5)$$

and the divergence between Q_x^n and the source P as

$$D(Q_x^n\|P) = \sum_{u\in\boldsymbol{X}} \sum_{\nu\in\boldsymbol{X}^l} q_x^n(u, \nu) \log \frac{q_x^n(u|\nu)}{P(u|\nu)}, \quad (2.6)$$

where $P(u|\nu)$ is the conditional probability of $x_i = u$ given that $s_i = \nu$.

The following lemma, which is a straightforward extension of a result due to Hoeffding [9], proposes an asymptotically optimal sequence of decision rules under the criterion M defined in Section I. This lemma will serve as a basic tool for examining the GLRT.

Lemma 1: The sequence of decision rules $\{\Lambda_n^*\}_{n\ge 1}$, where

$$\Lambda_n^* = \{x^n: D(Q_x^n\|P_0) \ge \lambda\}, \quad (2.7)$$

satisfies the constraint

$$\limsup_{n\to\infty} \frac{1}{n} \log P_0(\Lambda_n^*) \le -\lambda, \quad (2.8)$$

and at the same time maximizes, among all sequences of decision rules satisfying (1.4), the second kind error exponent $-\limsup_{n\to\infty} n^{-1} \log P_1(\overline{\Lambda}_n)$ for all $P_1 \in \boldsymbol{P}$.

Proof: To prove (2.8), first observe that the membership of a string x^n in Λ_n^* depends only on its empirical distribution (type). Hence,

$$P_0(\Lambda_n^*) = \sum_{T_x^n \subset \Lambda_n^*} P_0(T_x^n). \quad (2.9)$$

From [4, Lemma 1] we find that

$$\left|\frac{1}{n} \log P_0(T_x^n) + D(Q_x^n\|P_0)\right| \le \varepsilon_n, \quad (2.10)$$

where $\varepsilon_n = O(n^{-1} \log n)$ independently of x^n. Furthermore, the total number of distinct types T_x^n in $\boldsymbol{X}^n$ never exceeds $(n+1)^{X^{2l}}$ [12, p. 434], where X is the cardinality of the source alphabet $\boldsymbol{X}$. Combining this fact with (2.9) and (2.10), we have for every $\varepsilon > 0$,

$$\begin{aligned} P_0(\Lambda_n^*) &\le (n+1)^{X^{2l}} \max_{T_x^n \in \Lambda_n^*} P_0(T_x^n) \\ &\le (n+1)^{X^{2l}} \exp_2\left\{-n\left[\min_{T_x^n \subset \Lambda_n^*} D(Q_x^n\|P_0) - \varepsilon_n\right]\right\} \\ &= \exp_2\left\{-n\left[\lambda - \varepsilon_n - \frac{X^{2l}}{n} \log(n+1)\right]\right\} \le 2^{-n(\lambda-\varepsilon)}, \end{aligned} \quad (2.11)$$

where the equality follows from the definition of Λ_n^*. This completes the proof of (2.8).

As for the second part of the lemma, since P_1 is assumed an lth order Markov source, the empirical distribution Q_x^n serves as sufficient statistics for optimal decision under the criterion M (see also [3]–[7]). Let $\{\Lambda_n\}_{n\ge 1}$ be an arbitrary sequence of

decision rules satisfying (1.4), where membership in Λ_n depends only on the empirical distribution. Then, from (1.4), for every $\varepsilon > 0$, all large n and any $y^n \in \Lambda_n$,

$$2^{-(\lambda+\varepsilon_n)n} \geq P_0(\Lambda_n) \geq P_0(T_y^n) \geq \exp_2\{-n[D(Q_y^n\|P_0) + \varepsilon_n]\}, \quad (2.12)$$

where we have used again (2.10). This means that y^n is also in Λ_n^*, i.e., $\Lambda_n \subseteq \Lambda_n^*$, hence $P_1(\overline{\Lambda}_n^*) \leq P_1(\overline{\Lambda}_n)$ for every $P_1 \in \boldsymbol{P}$, completing the proof of Lemma 1. □

Remarks:

1) An important property of the test (2.7) is that it is *independent* of the class $\boldsymbol{P}$. Moreover, it is easily replaced by a test based on the LZ algorithm (see, e.g., [3]–[6], [8]).
2) Note that $D(Q_x^n\|P_0)$ in (2.7) can be rewritten as

$$D(Q_x^n\|P_0) = -H(Q_x^n) - \frac{1}{n}\log P_0(x^n), \quad (2.13)$$

where $-H(Q_x^n)$, in turn, can be also expressed as

$$-H(Q_x^n) = \frac{1}{n}\log P_1(x^n) + D(Q_x^n\|P_1). \quad (2.14)$$

Thus, if $\boldsymbol{P}$ is sufficiently rich so that

$$\inf_{P_1 \in \boldsymbol{P}} D(Q_x^n\|P_1) = 0, \qquad \forall x^n \in \boldsymbol{X}^n, \quad (2.15)$$

then it follows from (2.14) that $-H(Q_x^n) = n^{-1}\sup_{P_1 \in \boldsymbol{P}} \log P_1(x^n)$, in which case the GLRT (1.3) with $T_n(\lambda) = \lambda$ coincides with Hoeffding's test (2.7) and hence, it is asymptotically optimal. Indeed, in all cases considered in [3]–[6], l was assumed fixed and $\boldsymbol{P}$ was the entire class of lth-order Markov sources, therefore (2.15) holds trivially. It should be noted again that even if a general FS source can be approximated by a high-order Markov source, the latter might have many more states than that of the former, thus (2.15) may not hold.

In the next section, we demonstrate that even in simpler situations the GLRT is not always asymptotically optimal. Before we present such a counterexample, however, we need a preliminary step.

It can be shown, using a technique similar to the proof of Lemma 1, that for a given $P_1 \in \boldsymbol{P}$, the second kind error exponent, associated with Hoeffding's test (2.7), is given by

$$-\lim_{n\to\infty}\frac{1}{n}\log P_1(\overline{\Lambda}_n^*) = e(\lambda) \triangleq \inf_{Q\in A} D(Q\|P_1), \quad (2.16)$$

where A is the set of all lth-order Markov measures Q, defined as

$$A = \{Q: D(Q\|P_0) \leq \lambda\}. \quad (2.17)$$

The function $e(\lambda)$ is called the error-exponent function [13, p. 123, Definition 4.6.1]. Since B is a convex set and $D(\cdot\|P_1)$ is a strictly convex function, the infimum (2.16) is attained uniquely by a measure $\tilde{Q}$ on the boundary of A, i.e., $D(\tilde{Q}\|P_0) = \lambda$. Specifically, by standard Lagrange minimization (see also [13, ch. 4]), we find that $\tilde{Q}$ is of the form

$$\tilde{Q}(u|v) = K_v P_0^s(u|v)P_1^{1-s}(u|v), \qquad \forall u \in \boldsymbol{X}, \qquad v \in \boldsymbol{X}^l, \quad (2.18)$$

where $s \in [0,1]$ is chosen such that $D(\tilde{Q}\|P_0) = \lambda$ and $K_v = [\Sigma_{u\in\boldsymbol{X}} P_0^s(u|v)P_1^{1-s}(u|v)]^{-1}$ is a normalization factor. It is easy to see that the same measure $\tilde{Q}$ minimizes $D(Q\|P_0)$ over the set

$$B = \{Q: D(Q\|P_1) \leq e(\lambda)\}. \quad (2.19)$$

Consider now a sequence $\{T_x^n\}_{n\geq 1}$ of types with $Q_x^n \to Q_0$ for some lth order Markov measure Q_0 in the interior of B, i.e., $D(Q_0\|P_1) < e(\lambda)$, and suppose that for some sequence of decision rules $\{\Lambda_n\}_{n\geq 1}$, satisfying (1.4), T_x^n is classified into $P = P_0$ for infinitely many values of n. Then,

$$-\limsup_{n\to\infty}\frac{1}{n}\log P_1(\overline{\Lambda}_n) \leq -\limsup_{n\to\infty}\frac{1}{n}\log P_1(T_x^n) = D(Q_0\|P_1) < e(\lambda), \quad (2.20)$$

which means, in view of (2.16), that this sequence of decision rules is not optimal. We have just proved the following lemma, which provides a necessary condition for optimality of any given sequence of decision rules.

Lemma 2: Let $\{\Lambda_n\}_{n\geq 1}$ be an asymptotically optimal sequence of decision rules under the criterion M, where Q_x^n is sufficient statistics for Λ_n, and let $\{T_x^n\}_{n\geq 1}$ be a sequence of types such that the corresponding sequence of empirical distributions $\{Q_x^n\}_{n\geq 1}$ has an accumulation point Q_0 in the interior of B for some $P_1 \in \boldsymbol{P}$. Then, for all large n, $T_x^n \subseteq \Lambda_n$.

In other words, the lemma states that every string x^n whose empirical distribution falls in a "sphere" B with radius $e(\lambda)$, centered at P_1, must be classified into $P = P_1$ for sufficiently large n. As an example of an asymptotically optimal sequence of decision rules, other than Hoeffding's test, suppose that P_1 is known (i.e., $\boldsymbol{P} = \{P_1\}$ and consider the LRT (1.2) with the optimal threshold function. Since both Hoeffding's test and the LRT are asymptotically optimal, in this case, then by comparing the performance of the two tests, it is easy to show that the threshold function $T_n(\lambda)$ of the optimal LRT tends to a limit $T(\lambda)$ given by

$$T(\lambda) = \lambda - e(\lambda). \quad (2.21)$$

This relation can be also obtained from a straightforward extension of [13, p. 120, Theorem 4.5.2].

III. A Counterexample

Let P_0 be a binary memoryless (Bernoulli) source with letter probabilities $P_0(\text{"0"}) = P_0(\text{"1"}) = 0.5$. Let $\boldsymbol{P}$ be the class of all Bernoulli sources P_1 where the parameter $\theta \triangleq P_1(\text{"0"})$ is either larger than 0.7 or smaller than 0.2. The idea behind this choice of an asymmetric $\boldsymbol{P}$, is that the "distances" of its two parts, $\{\theta: \theta \leq 0.2\}$ and $\{\theta: \theta \geq 0.7\}$, from $\theta = 0.5$, which corresponds to P_0, are different and hence a fixed threshold function $T_n(\lambda)$ cannot fit both parts of $\boldsymbol{P}$. Throughout this section, we shall denote a Bernoulli source with parameter θ, $0 \leq \theta \leq 1$, by P^θ. Since all sources involved are memoryless, the relative frequencies of letters $q_x^n(u) \triangleq n^{-1}\Sigma_{i=1}^n \delta(x_i, u)$, $u =$ "0", "1", are sufficient statistics in this case.

We now show that for $\lambda = D(P^{0.4}\|P^{0.5}) = D(P^{0.6}\|P^{0.5})$ there is no threshold value $T_n(\lambda)$ for which the GLRT is asymptotically optimal. Define

$$\lambda_1 = D(P^{0.6}\|P^{0.7}), \quad (3.1a)$$

$$\lambda_2 = D(P^{0.4}\|P^{0.2}). \quad (3.1b)$$

Assume first, that $T_n(\lambda) \leq \lambda - \lambda_1$ and consider the type of strings x^n for which $q_x^n(\text{"0"}) = 0.6$. For the GLRT to satisfy the constraint (1.4), this type must be classified into P_0. Since

$\sup_{P_1 \in \boldsymbol{P}} P_1(x^n) = P^{0.7}(x^n)$, as can be easily shown, we have that

$$\begin{aligned}\frac{1}{n}\log\sup_{P_1\in\boldsymbol{P}} P_1(x^n) - \frac{1}{n}\log P_0(x^n) &= \frac{1}{n}\log P^{0.7}(x^n) - \frac{1}{n}\log P_0(x^n)\\ &= D(P^{0.6}\|P^{0.5}) - D(P^{0.6}\|P^{0.7}) = \lambda - \lambda_1 \ge T_n(\lambda),\end{aligned} \tag{3.2}$$

which means that the GLRT classifies x^n into $P = P_1$ and hence contradicts the previous requirement that x^n should be classified to P_0.

Assume next, that $T_n(\lambda) > \lambda - \lambda_1$, and consider a string x^n for which $q_x^n(\text{"0"}) = 0.35$. Since $D(P^{0.35}\|P_0) > \lambda$ and $P^{0.35}$ lies "between" $P_1 = P^{0.2} \in \boldsymbol{P}_1$ and $P_0 = P^{0.5}$, then $D(P^{0.35}\|P^{0.2}) < e(\lambda)$, and hence by Lemma 2, x^n must be classified into $P = P_1$. Since $D(P^{0.35}\|P^{0.2}) < D(P^{0.35}\|P^{0.7})$, this implies that $\sup_{P_1\in\boldsymbol{P}} P_1(x^n) = P^{0.2}(x^n)$. Finally, since

$$D(P^{0.35}\|P^{0.5}) - D(P^{0.35}\|P^{0.2}) < \lambda - \lambda_1,$$

as can one easily verify, we find that

$$\begin{aligned}\frac{1}{n}\log\sup_{P_1\in\boldsymbol{P}} P_1(x^n) - \frac{1}{n}\log P_0(x^n) &= D(P^{0.35}\|P^{0.5}) - D(P^{0.35}\|P^{0.2}) < \lambda - \lambda_1 < T_n(\lambda),\end{aligned} \tag{3.3}$$

which implies that the GLRT classifies x^n into P_0 in contrast to the previous requirement that x^n should be classified into P_1.

Since we have contradicted both possibilities $T_n(\lambda) \le \lambda - \lambda_1$ and $T_n(\lambda) > \lambda - \lambda_1$, we have proved the nonexistence of a threshold value for which the GLRT is asymptotically optimal under the criterion M.

IV. Conditions for Asymptotic Optimality of the GLRT

We begin by deriving a necessary and sufficient condition for the asymptotic optimality of the GLRT in the situation described in Section II. For the sake of simplicity and in view of the Remark 2 in Section II, we confine attention to the case $T_n(\lambda) = \lambda$.

Theorem 1: Let $\boldsymbol{P} \cup \{P_0\}$ be a subset of the class of stationary, ergodic lth-order Markov sources. Then, the GLRT with a threshold function $T_n(\lambda) = \lambda$ is asymptotically optimal under criterion M, if and only if for all $P_1 \in \boldsymbol{P}$,

$$\inf_{Q\in C} D(Q\|P_1) \ge e(\lambda), \tag{4.1}$$

where

$$C = \left\{Q\colon D(Q\|P_0) - \inf_{P_1\in\boldsymbol{P}} D(Q\|P_1) < \lambda \le D(Q\|P_0)\right\}. \tag{4.2}$$

The interpretation of the theorem is as follows. The right-hand side of (4.1) is the second kind error exponent associated with Hoeffding's test (2.7), which is always optimal by Lemma 1, and we want the second kind error exponent of the GLRT to be as large. The right-hand side of (4.1) corresponds to the exponential rate of the probability of the set $\Lambda_n^* - \Lambda_n^{\mathrm{GLRT}}$, namely, the set of all sequences which are classified to P_0 by the GLRT but not by Hoeffding's test. If this exponential rate is larger than Hoeffding's exponent, then the contribution of this difference set is exponentially negligible and the two tests provide the same performance. Note that if the simple sufficient condition (2.15) is satisfied, as in the case where $\boldsymbol{P}$ contains all Markov sources of order l or less, then (4.1) holds trivially because C becomes an empty set.

Proof of Theorem 1: We first show that the GLRT with $T_n(\lambda) = \lambda$ satisfies (1.4). To this end, it is sufficient to show that $\Lambda_n^{\mathrm{GLRT}} \subseteq \Lambda_n^*$, as we already know from Lemma 1 that Λ_n^* satisfies (1.4). This, in turn, follows from the following consideration:

$$\begin{aligned}\frac{1}{n}\log\sup_{P_1\in\boldsymbol{P}} P_1(x^n) - \frac{1}{n}\log P_0(x^n) &= -H(Q_x^n) - \inf_{P_1\in\boldsymbol{P}} D(Q_x^n\|P_1) - \frac{1}{n}\log P_0(x^n)\\ &\le -H(Q_x^n) - \frac{1}{n}\log P_0(x^n)\\ &= D(Q_x^n\|P_0),\end{aligned} \tag{4.3}$$

where the inequality follows from the fact that a divergence is nonnegative. Thus, (4.3) tells us that if x^n is such that the left-most side exceeds λ, namely, $x^n \in \Lambda_n^{\mathrm{GLRT}}$, then the right-most side exceeds λ as well, i.e., $x^n \in \Lambda_n^*$. This completes the proof that $\Lambda_n^{\mathrm{GLRT}} \subseteq \Lambda_n^*$ or, equivalently, $\overline{\Lambda}_n^* \subseteq \overline{\Lambda}_n^{\mathrm{GLRT}}$.

In view of this fact, in order to examine $P_1(\overline{\Lambda}_n^{GLRT})$, we can decompose this probability as

$$P_1(\overline{\Lambda}_n^{\mathrm{GLRT}}) = P_1(\overline{\Lambda}_n^*) + P_1(\overline{\Lambda}_n^{\mathrm{GLRT}} \cap \Lambda_n^*). \tag{4.4}$$

The first term on the right-hand side is the second kind error probability associated with Hoeffding's test, which decays exponentially with rate $e(\lambda)$, as mentioned earlier. Similarly, the second term on the right-hand side of (4.4) decays with an exponential rate given by the left-hand side of (4.1), as all sequences in $\overline{\Lambda}_n^{\mathrm{GLRT}} \cap \Lambda_n^*$ have their empirical distributions in C. Therefore, we observe that the GLRT is asymptotically optimal, if and only if the exponential rate of the second term of (4.4) is at least as large as that of the first term, which is exactly condition (4.1). This completes the proof of Theorem 1. □

The necessary and sufficient condition of Theorem 1 is difficult to check in general. We, therefore, derive a sufficient condition (Theorem 2) for the asymptotic optimality of the GLRT, which is easier to verify.

Theorem 2: Let $\boldsymbol{P} \cup \{P_0\}$ be a subset of the class of stationary, ergodic lth-order Markov sources, and let $\tilde{Q} = \arg\min_{Q\in B} D(Q\|P_0)$, where B is defined as in (2.19). If $\tilde{Q} \in \boldsymbol{P}$ for all $P_1 \in \boldsymbol{P}$, then the GLRT with the threshold function $T_n(\lambda) = \lambda$ is asymptotically optimal under criterion M.

Since $\tilde{Q}$ is given by an "exponential combination" of P_0 and P_1 (see (2.18)), the significance of Theorem 2 is that if $\boldsymbol{P} \cup \{P_0\}$ is closed with respect to such exponential combinations, then the GLRT is asymptotically optimal. It is easy to see that this closure property holds if $\boldsymbol{P}$ is an exponential family of measures as is the case in each of the earlier studies [3]–[7]. This holds true even if one uses statistics of order higher than that of the sufficient statistics (e.g., when $\boldsymbol{P}$ is a subclass of the first-order

Markov processes but Q_x^n corresponds to $l > 1$). In this case, (2.15) is not satisfied, yet $\tilde{Q} \in \boldsymbol{P}$ and hence, the GLRT is asymptotically optimal. This condition holds also for many other subclasses $\boldsymbol{P}$ of the lth-order Markov sources, e.g., the subclass defined by dependence only upon even lags of the process, i.e.,

$$P(x_i|x^{i-1}) = P(x_i|x_{i-2}, x_{i-4}, \cdots, x_{i-l}), \tag{4.5}$$

where l is assumed even. Again, this class does not satisfy the simple sufficient condition (2.15) because there exist empirical lth-order Markov measures Q_x^n that are far apart from any Markov measure in the subclass defined by (4.5), e.g., measures Q_x^n for which x_i depends strongly on odd lags x_{i-1}, x_{i-3}, etc. For binary memoryless sources, Theorem 2 means convexity of $\boldsymbol{P} \cup \{P_0\}$, which is clearly not the case in the counterexample of Section III. Finally, we point out that since Theorem 2 provides merely a sufficient condition, it is easy to find an example where this condition does not hold and yet the GLRT is asymptotically optimal. A simple example is when $\boldsymbol{P}$ includes only one measure P_1, in which case the GLRT coincides with the LRT and hence optimal, however, the condition is clearly not satisfied.

Proof of Theorem 2: Since by assumption $\tilde{Q} \in \boldsymbol{P}$, then,

$$D(Q\|\tilde{Q}) \geq \inf_{P_1 \in \boldsymbol{P}} D(Q\|P_1), \tag{4.6}$$

and hence,

$$\inf_{Q \in C} D(Q\|P_1) \geq \inf_{Q \in D} D(Q\|P_1), \tag{4.7}$$

where

$$D = \{Q\colon D(Q\|P_0) - D(Q\|\tilde{Q}) < \lambda \leq D(Q\|P_0)\}. \tag{4.8}$$

Thus, to prove that (4.1) holds, and hence the GLRT is asymptotically optimal, it is sufficient to show that

$$\inf_{Q \in D} D(Q\|P_1) \geq \inf_{Q \in A} D(Q\|P_1) = e(\lambda). \tag{4.9}$$

Conversely, assume that (4.9) is false, namely, the measure $\hat{Q} \in D$ that minimizes $D(Q\|P_1)$ on the left-hand side of (4.10) is an interior point of B. Then, it follows from [14, Theorem 2.2] that

$$D(\hat{Q}\|P_0) - D(\hat{Q}\|\tilde{Q}) \geq D(\tilde{Q}\|P_0) = \lambda, \tag{4.10}$$

which contradicts the assumption that $\hat{Q} \in D$ and hence, completes the proof of Theorem 2. □

Remarks:

1) M. Gutman has shown us that a condition somewhat different than that of Theorem 2 is actually both necessary and sufficient. Specifically, referring to the notations of Theorem 2, the GLRT is asymptotically optimal, if and only if

$$\forall P_1 \in \boldsymbol{P}, \qquad \inf_{P' \in \boldsymbol{P}} D(\tilde{Q}\|P') = 0.$$

2) Although it has been assumed in this correspondence that P_0 is completely known, the results can be easily generalized to the case where P_0 is unknown except that it is in a subset $\boldsymbol{Q}$ of the lth-order Markov sources. In this case, (1.4) is required to hold for all $P_0 \in \boldsymbol{Q}$, or, equivalently,

$$\limsup_{n \to \infty} \frac{1}{n} \log \sup_{P_0 \in \boldsymbol{Q}} P_0(\Lambda_n) < -\lambda. \tag{4.11}$$

Similarly, $P_0(x^n)$ in the GLRT (1.3) will be replaced by $\sup_{P_0 \in \boldsymbol{Q}} P_0(x^n)$. In all cases studied in [3]–[8], $\boldsymbol{Q}$ was a subset of $\boldsymbol{P}$.

3) We conjecture that for the case where $\boldsymbol{P}$ is the class of all finite-state sources with a given number of states (see Appendix and [8]), the condition of Theorem 2, and hence also the asymptotic optimality of the GLRT, holds true. Although we have not been able to prove it analytically, a computer search over the space $\boldsymbol{P}$ of two-state binary sources, with $l = 7$ and P_0 being a binary memoryless symmetric source (similar to that of Section III), failed to provide a counterexample for which $\tilde{Q}$, as defined in (2.18), falls outside $\boldsymbol{P}$.

Appendix
Approximation of FS Sources by Unifilar FS Sources

An FS source P, (also called a hidden Markov source [15],) is characterized by an observed output process $(x_1, x_2, \cdots)$, $x_i \in \boldsymbol{X}$, and a corresponding unobserved (hidden) finite-alphabet state process $(s_1, s_2, \cdots)$, $s_i \in \boldsymbol{S}$, which are jointly Markov, i.e.,

$$P(x^n, s^{n+1}) = \prod_{i=1}^{n} p(x_i, s_{i+1}|s_i), \tag{A.1}$$

where we shall assume that the initial state $s_1 = \sigma$ is fixed. If, in addition,

$$s_{i+1} = g(x_i, s_i) \tag{A.2}$$

with probability 1, for some mapping $g\colon \boldsymbol{X} \times \boldsymbol{S} \to \boldsymbol{S}$, then (A.1) is referred to as an *unifilar* FS source. In this case, the state process can be reconstructed recursively from the observation process using (A.2). An important special case of a unifilar FS source is an lth-order Markov Source where $s_i = x_{i-l}^{i-1}$.

We show that a general FS source (A.1) with strictly positive transition probabilities $\{p(u, \nu|w)\}_{u \in \boldsymbol{X}, \nu, w \in \boldsymbol{S}}$ can be approximated by a unifilar FS source with a particular choice of g and sufficiently many states. Suppose that $p(u, \nu|w) \geq \delta > 0$ for all $u \in \boldsymbol{X}$, $\nu, w \in \boldsymbol{S}$, and hence,

$$\delta \cdot p(u, \nu|\sigma) \leq p(u, \nu|w) \leq \delta^{-1} p(u, \nu|\sigma).$$

Assume further that l divides n and parse x^n into n/l nonoverlapping blocks of length l. Then,

$$\begin{aligned}
P(x^n) &= \sum_{s_2^{n+1}} \prod_{i=1}^{n} p(x_i, s_{i+1}|s_i) \\
&= \sum_{s_2^{n+1}} \prod_{j=0}^{n/l-1} \prod_{i=1}^{l} p(x_{jl+i}, s_{jl+i+1}|s_{jl+i}) \\
&\geq \sum_{s_2^{n+1}} \prod_{j=0}^{n/l-1} \delta \cdot p(x_{jl+1}, s_{jl+2}|\sigma) \\
&\quad \cdot \prod_{i=2}^{l} p(x_{jl+i}, s_{jl+i+1}|s_{jl+i}) \\
&= \delta^{n/l} \prod_{j=0}^{n/l-1} \sum_{s_{jl+2}^{jl+l+1}} p(x_{jl+1}, s_{jl+2}|\sigma) \\
&\quad \cdot \prod_{i=2}^{l} p(x_{jl+i}, s_{jl+i+1}|s_{jl+i})
\end{aligned}$$

$$= 2^{-nl^{-1}\log(1/\delta)} \prod_{j=0}^{n/l-1} P\left(x_{jl+1}^{jl+l}\right). \quad \text{(A.3)}$$

In a similar manner we obtain

$$P(x^n) \le 2^{nl^{-1}\log(1/\delta)} \prod_{j=0}^{n/l-1} P\left(x_{jl+1}^{jl+l}\right), \quad \text{(A.4)}$$

which together with (A.3) means that (A.1) can be approximated by a block memoryless source of l-tuples (see also [16, Appendix]) defined by

$$\tilde{P}(x^n) = \prod_{j=0}^{n/l-1} P\left(x_{jl+1}^{jl+l}\right) = \prod_{j=0}^{n/l-1} \prod_{i=1}^{l} P\left(x_{jl+i} | x_{jl+1}^{jl+i-1}\right). \quad \text{(A.5)}$$

Under the measure $\tilde{P}$, x_i depends at most on the l preceding letters x_{i-l}^{i-1} and the induced conditional probability $P(x_i|x_{i-l}^{i-1})$, in turn, depends on the position of i with respect to the block endpoints, i.e., on i mod l. Hence, (A.5) can be described as a unifilar FS source with no more than $l \cdot X^l$ states (X is the alphabet size), where the state variable $\tilde{s}_i$ consists of the l preceding letters and a modulo-l time counter c_i, i.e., $\tilde{s}_i = (x_{i-l}^{i-1}, c_i)$. Thus, from (A.3) and (A.4) we find that if $l > \varepsilon^{-1}\log(1/\delta)$, then with probability 1,

$$2^{-n\varepsilon} \prod_{i=1}^{n} P(x_i|\tilde{s}_i) \le P(x^n) \le 2^{n\varepsilon} \prod_{i=1}^{n} P(x_i|\tilde{s}_i), \quad \text{(A.6)}$$

which is similar to (2.2) with the generalization that x_{i-l}^{i-1} is replaced by $\tilde{s}_i$.

References

[1] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, vol. 20-A, pp. 175–240, 264–299, 1928.

[2] H. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.

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

[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] ——, "On tests for independence, tests for randomness and universal data compression," submitted for publication.

[6] 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.

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

[8] 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.

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

[10] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[11] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, pp. 285–290, Mar. 1991; correction, vol. 37, p. 698, May 1991.

[12] L. D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 431–438, July 1981.

[13] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[14] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[16] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

Recursively Indexed Quantization of Memoryless Sources

Khalid Sayood, *Member, IEEE*, and Sangsin Na, *Member, IEEE*

***Abstract*—A recursively indexed scalar quantizer that performs as well as high-dimensional vector quantizers for several important sources, without the attendant complexity is presented.**

***Index Terms*—Quantization, source coding.**

I. Introduction

Consider the source coding scenario of Fig. 1. A k-dimensional N-point (or output vector) quantizer Q is a mapping of the k-dimensional Euclidean space $\boldsymbol{R}^k$ to a set of k-dimensional quantization vectors $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_{N-1}\}$. The quantization rule, denoted by Q, is

$$Q(\boldsymbol{x}) = \boldsymbol{y}_j, \qquad \text{if } \boldsymbol{x} \in S_j,$$

where $\{S_j: j = 0, 1, \cdots, N-1\}$ is a partition of $\boldsymbol{R}^k$, and $\{S_j\}$ are called the quantization regions. A fixed-to-fixed length binary encoder is a mapping from the set $\{0, 1, \cdots, N-1\}$ of the indices of the quantization vectors to a set of binary sequences of length $\lceil \log_2 N \rceil$, where $\lceil x \rceil$ is the smallest integer not exceeded by x. A fixed-to-variable length binary encoder is a mapping from the set $\{0, 1, \cdots, N-1\}$ of the indices of the quantization vectors to a set of binary sequences of varying lengths. The quantizer operates in the following fashion: in the first operation cycle the quantizer Q takes the first k source symbols $X_1, X_2, \cdots, X_k$ and produces the corresponding reproduction symbols $Y_1, Y_2, \cdots, Y_k$. It then enters into the next operation cycle, where it takes the next k source symbols $X_{k+1}, X_{k+2}, \cdots, X_{2k}$ to produce $Y_{k+1}, Y_{k+2}, \cdots, Y_{2k}$. The binary encoder takes a block of indices of reproduction vectors and encodes it into a binary sequence for transmission over a noiseless channel or for storage. At the receiver, the received binary sequence is decoded into a sequence of reproduction indices, which are mapped back to the reproduction vectors via stored quantization vectors.

The performance of a source coding scheme is often measured by its rate and distortion. The rate is the average number of binary digits used for representing a source symbol. When a fixed-to-fixed length binary encoder is used, the rate is $(1/k)\lceil \log_2 N \rceil$ where N is the size of the reproduction alphabet and k is the dimension of the input vectors. The distortion is in some sense a measure of the average closeness of the reproduction sequence to the source sequence. In this correspondence, we will measure the distortion by $(1/k)E\{\|\boldsymbol{X} - Q(\boldsymbol{X})\|^2\}$, where the expectation E is taken with respect to the k-dimensional source distribution $p(x)$ and where $\|\cdot\|$ denotes the L^2 norm.

Manuscript received January 31, 1991; revised January 17, 1992. This work was supported by the NASA Lewis Research Center under Grant NAG3-806. This work was presented in part at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 1990.

K. Sayood is with the Department of Electrical Engineering and the Center for Communication and Information Science, University of Nebraska, 209 N. WSEC, Lincoln, NE 68588-0511.

S. Na is with the Department of Electronic Engineering, Ajou University, 5 Wonchun-Dong, Suwon, Korea.

IEEE Log Number 9200887.