SRC TR 89-25

MARYLAND

TECHNICAL RESEARCH REPORT

Optimal And Robust Memoryless Discrimination From Dependent Observations

By

D. Sauder
and
E. Geraniotis

SYSTEMS RESEARCH CENTER

UNIVERSITY OF MARYLAND

COLLEGE PARK, MARYLAND 20742

OPTIMAL AND ROBUST MEMORYLESS DISCRIMINATION FROM DEPENDENT OBSERVATIONS

Douglas Sauder and Evaggelos Geraniotis

Department of Electrical Engineering
and Systems Research Center
University of Maryland
College Park, MD 20742

ABSTRACT

In this paper we consider discrimination between two possible sources based on observations of their output. The discrimination problem is modeled by means of a general binary hypothesis test, the main emphasis being on situations that cannot be modeled as signals in additive noise. The structure of the discriminator is such that the observations are passed through a memoryless nonlinearity and summed up to form a test statistic, which is then compared to a threshold. In this paper we consider only fixed sample size tests. Four different performance measures, which resemble the signal-to-noise ratios encountered in the signal in additive noise problems, are derived under different problem formulations. The optimal nonlinearities for each of the performance measures are derived as solutions to various integral equations. For three of the four performance measures, we have successfully obtained robust nonlinearities for uncertainty in the marginal and the joint probability density functions of the observations. Computer simulation results which demonstrate the advantage of using our nonlinearities over the i.i.d. nonlinearity under the probability of error criterion are presented.

This research was supported in part by the Naval Research Laboratory/Systems Research Center Fellowship Program at the University of Maryland, College Park, through ONR Grant N00014-85-G-0213.

I. Introduction

In this paper, we have chosen to use the term "discrimination" to refer to a detection problem in which the "noise" characteristics under the two hypotheses are substantially different. This is in contrast to the case of an additive noise channel, where it is assumed that an independent noise process is added to the signal as it traverses the channel. In our model, we may assume that all the randomness is in the signal itself, and that transmission through the channel is noiseless. This model is especially appropriate in the context of radar in problems sometimes referred to as target discrimination or target identification. In such problems, the output of the signal processing device must indicate which of several targets is present. This discrimination or identification is to be performed after the initial decision that some object is present. Thus under each possible hypothesis, one observes the random output of a particular source and must decide which particular source is present. For the purposes of our paper, we shall state the problem as a binary hypothesis testing problem:

$$\begin{aligned} H_0 &: \quad \{X_k\} \text{ has the distribution } F_0 \\ H_1 &: \quad \{X_k\} \text{ has the distribution } F_1. \end{aligned} \tag{1.1}$$

The observation process $\{X_k\}$ is assumed to be stationary and strong mixing (which implies ergodicity) and we denote by f_i, $f_i^{(n)}$ the marginal and nth-order joint densities of the process with respect to the measure ν under H_i. Throughout this paper, as in (1.1) above, we use the symbol F_i to denote the distribution of the entire process under H_i, which in general will not be iid. We will be concerned with the asymptotic performance of various discriminators, but we assume that the process distributions remain fixed, which is in contrast to the ARE performance criterion where the "distance" between the two distributions converges to zero as the sample size increases.

It is the goal of this paper to present a thorough discussion of the use of memoryless fixed sample size (FSS) methods to discriminate between two sources which produce correlated outputs.

These memoryless discriminators are characterized by their use of a test statistic of the form

$$T_n(\mathbf{x}) = \sum_{i=1}^{n} g(x_i) \tag{1.2}$$

where g is a Borel measurable real-valued function and where $\mathbf{x} = (x_1, \ldots, x_n)$ represents the observation vector of size n. Our discriminators have the form

$$\text{decide } H_1 \quad \text{iff} \quad T_n(\mathbf{x}) \in \Gamma \subset \mathbf{R}. \tag{1.3}$$

If $\mathrm{E}_0\, g(X_1) < \mathrm{E}_1\, g(X_1)$ and n is sufficiently large, then we lose very little in terms of error probabilities if we take Γ to be an interval $[n\gamma, \infty)$ defined by a threshold $n\gamma$, for if the processes are ergodic, then $\frac{1}{n}T_n \to \mathrm{E}_i\, g(X_1)$ almost surely under H_i as $n \to \infty$.

The most obvious advantage of such a discriminator is the simplicity of its implementation. Indeed, the structure of the test statistic T_n suggests that it might be suitable for sequential discrimination. Furthermore, we shall see that asymptotically optimal test statistics of the form (1.2) can be calculated from only the marginal and bivariate joint densities of the processes, and thus they provide an alternative to the likelihood ratio in cases where the nth-order densities are unknown or lack closed form expressions. In the case where iid observations are produced under both hypotheses, the LRT has the form of a memoryless discriminator with $g = \log(f_1/f_0)$, and is therefore optimal under any probability of error criterion. On the other hand, if the situation is such that there is correlation of the observations under at least one of the hypotheses, then LRT will require memory, and consequently, any memoryless discriminator will be suboptimal.

Situations in which the LRT does not have a closed form, such as in many cases involving nonlinear transformations of Gaussian processes, are commonly encountered. The discriminators which we present here will provide alternatives in such situations, and show promise of significant improvement over the iid nonlinearity $g = \log(f_1/f_0)$.

Define the functionals μ_i and $\sigma_i \geq 0$ by

$$\mu_i(g) = \mathrm{E}_i\, g(X_1) \tag{1.4}$$

$$\sigma_i^2(g) = \mathrm{Var}_i\, g(X_1) + 2\sum_{j=1}^{\infty} \mathrm{Cov}_i[g(X_1), g(X_{j+1})]. \tag{1.5}$$

where E_i, Var_i, and Cov_i denote, respectively, th expectation, variance and covariance operators performed under H_i. If $\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}_i\, T_n$ exists, then it is precisely σ_i^2 as defined above. For our results, we must assume that the normalized test statistic $(T_n - n\mu_i)/\sqrt{n\sigma_i^2}$ converges in distribution to a standard normal random variable. Conditions on strong mixing processes which imply convergence to the normal distribution can be found in the survey papers [19], [20] or in some of the other papers on memoryless detection methods such as [2]. See also the appendix of this paper.

Since the error probabilities are intractable for a discriminator of the form (1.3), we proceed in our analysis by considering performance measures which resemble signal-to-noise ratios. A fundamental difficulty which arises here is determining a quantity which represents the "noise," since there is no additive noise involved in our model. Earlier work in the area of memoryless detection has focused primarly on the detection of a weak signal in noise, where the noise process is additive and is independent of the signal. In such cases, the "noise" power is determined to be the variance of the test statistic, and there is no ambiguity in such a definition because in the limit as the sample size increases, the variances under the two hypotheses are equal. Therefore, the efficacy performance measure, which resembles a signal-to-noise ratio with the noise power represented in this way, has been thoroughly studied and has led to useful results ([1], [2]). For the discrimination problem, finding a useful measure of the noise power is a nontrivial problem, since the variances of the test statistic under the two hypotheses cannot be assumed to be equal.

We derive in this paper four different performance measures, including a new performance measure S_3 which has rather nice properties which make it mathematically tractable. Sadowsky and Bucklew [3] have recently derived a new performance measure similar to the efficacy functional in their work in nonlocal detection. Their performance measure is precisely the performance measure which we call S_2 in this paper. The performance measure which we call S_0 has been referred

to in the literature as the deflection [21], [16]. In all cases, we derive integral equations which can be solved to yield the optimal nonlinearity for that performance measure. Certain of these integral equations have been previously derived for the case of m-dependent processes. We show how to extend these results to the case of ρ-mixing processes. To conclude our results for this paper consider the problem of minimax robustness for three of the four performance measures, and determine the optimal robust nonlinearity for each case.

The paper is organized in the following way. In Section II, we consider performance measures S_0 and S_1, which are derived under asymptotic Neyman-Pearson formulations. The optimal nonlinearity in either case is shown to solve a linear integral equation when the pocesses are m-dependent, and a discussion is given concerning the solution of the integral equation. Finally, we show how to extend the results to the case of ρ-mixing processes. The organization of Section III is similar to that of Section II, but the discussion centers on the performance measures S_2 and S_3, which are derived from Chernoff bounds. The optimal nonlinearities are also given by the solution of integral equations, the integral equation for S_2 being nonlinear. We propose a scheme for solving the nonlinear integral equation by iteratively solving a series of linear integral equations. By an argument similar to that in Section II, we show how to extend the results to the case of ρ-mixing processes. In Section IV, the problem of robustness is formulated by specifying uncertainty classes for the two hypotheses. The least favorable distributions and the corresponding optimal robust nonlinearities are then derived for the performance measures S_0, S_1, and S_3. Section V contains numerical results which illustrate in a practical way the usefulness of our results. Receiver operating characteristic (ROC) curves are presented, and a comparison among the different discriminators is made based on the simulated error probabilities. The ROC for the optimal iid discriminator, when the input is correlated, is also given. Section VI is the concluding section and contains a practical discussion of some matters related to the implementation of such discriminators.

II. The Performance Measures S_0 and S_1

In this section, we examine the technique of memoryless discrimination under an asymptotic Neyman-Pearson formulation. We denote by P_i the probability of error when H_i is true, and we regard this quantity to be a function of the sample size n. For consistent tests, $P_i \to 0$ as $n \to \infty$. We will first consider the problem of maximizing the convergence rate of P_1 under the constraint that $P_0 \leq \alpha$. In the following subsection, we derive a performance measure S_1 which determines approximately this convergence rate. If we maximize instead the convergence rate of P_0 under the constraint that $P_1 \leq \alpha$ then we can derive in a similar way the performance measure S_0. We will state and prove results only for S_1.

A. The derivation of the performance measure S_1

The normalized test statistic is assumed to have a distribution which is approximately Gaussian when n is large, as discussed above. Without loss of generality, we may assume that $\mu_1 > \mu_0$, since the direction of the inequality depends on the sign of g. We may also exclude the case of $\mu_1 = \mu_0$ since this occurs only when the marginal densities are identical, and in such cases memoryless discrimination is difficult. For a test of the form (1.3) with Γ defined by a single threshold $\Gamma = [n\gamma, \infty)$, the error probabilities have the approximate values

$$\begin{aligned} P_0 &= \Phi\left[-\sqrt{n}\frac{\gamma - \mu_0}{\sigma_0}\right] \\ P_1 &= \Phi\left[\sqrt{n}\frac{\gamma - \mu_1}{\sigma_1}\right] \end{aligned} \tag{2.1}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Now in order to have $P_0 = \alpha$, we must take $\gamma = (\sigma_0/\sqrt{n})\Phi^{-1}(\alpha) + \mu_0$, and substituting for γ in the expression for P_1 we obtain

$$P_1 = \Phi\left[\frac{\sigma_0}{\sigma_1}\Phi^{-1}(\alpha) - \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_1}\right]. \tag{2.2}$$

Since $\Phi(x)$ is an increasing function of x, to minimize P_1 it is necessary to make the argument of Φ in (2.2) as small as possible; that is, to make the argument large in magnitude and negative. The term $-\sqrt{n}(\mu_1 - \mu_0)/\sigma_1$ is negative, since we are assuming that $\mu_1 > \mu_0$, and it increases in magnitude as n increases, so that $P_1 \to 0$. The quantity $(\mu_1 - \mu_0)/\sigma_1$ determines approximately the rate at which P_1 goes to zero, and we can see that the best asymptotic performance results when this quantity is maximized. We define the following peformance measure

$$S_1 = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2}. \tag{2.3}$$

If we consider a similar problem with the roles of P_0 and P_1 reversed, then we obtain by the same method as that above the performance measure

$$S_0 = \frac{(\mu_1 - \mu_0)^2}{\sigma_0^2}. \tag{2.4}$$

Because all the results for S_1 can be applied in a straightforword way to the performance measure S_0, we will not consider S_0 again until Section V, when numerical examples are presented. One might notice the similarity of S_1 (or S_0) to the efficacy performance measure. Indeed, both are asymptotic performance measures based on central limit theory. There is an important difference, however, in that the efficacy is a performance measure for local optimality (as in the case of a weak signal assumption, for example) whereas S_1 is a nonlocal performance measure. Nevertheless, because of the similarity the results in the next two subsections are quite similar to the results in [1] and [2].

B. The optimal nonlinearity for S_1

We consider the following optimization problem

$$\text{maximize} \qquad S_1(g) = \frac{\left[\mu_1(g) - \mu_0(g)\right]^2}{\sigma_1^2(g)} \tag{2.5}$$

subject to the constraints that $\mathrm{E}_i g^2(X_1) < \infty$ for $i = 0, 1$, and where μ_i and σ_1^2 are given by (1.4) and (1.5). In the next subsection conditions are given which guarantee that the opimal nonlinearity

g_1 derived in this section satisfies the constraint $\mathrm{E}_1 g_1^2(X_1) < \infty$. In order to have $\mathrm{E}_0 g_1^2(X_1) < \infty$, then, it is sufficient to require that $f_0(x)/f_1(x)$ be bounded for all x, and we shall make this assumption. We shall also take this condition to mean that $f_1(x) = 0 \Rightarrow f_0(x) = 0$ as well. In this subsection we shall require the rather stringent condition that the observed process be m-dependent under either hypothesis. Observe that for an m-dependent process, the expression for $\sigma_i^2(g)$ as given by (1.5) becomes

$$\sigma_i^2(g) = \mathrm{E}_i g^2(X_1) + \sum_{j=1}^{m} \mathrm{E}_i g(X_1) g(X_{j+1}) - (2m+1)\left[\mathrm{E}_i g(X_1)\right]^2 . \tag{2.6}$$

At the end of this section, we shall discuss the extension of the m-dependent results to the more general case of strong mixing processes, and show that for all practical purposes, strong mixing processes can be approximated by m-dependent ones.

The method for solving the optimization problem (2.5) is essentially the same as that in [1], with minor modifications. This techinique is also used for the optimization of the other performance measures. The optimal nonlinearity which solves the problem (2.5) is given by the solution of the following integral equation

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{f_1(x)} - 2\lambda \int K_1(x,y) g(y) \nu(dy) \tag{2.7}$$

with the kernel K_1 given by

$$K_1(x,y) = \frac{1}{f_1(x)} \sum_{j=1}^{m} \left[f_1^j(x,y) + f_1^j(y,x) \right] - (2m+1) f_1(y). \tag{2.8}$$

The parameter λ determines the scaling (and the sign) of g. This scaling is irrelevant to the performance of the discriminator; however, $\lambda > 0$ is a necessary condition for $\mu_1 > \mu_0$. In the analysis that follows, we show that a useful choice is $\lambda = \frac{1}{2}$. We will denote by g_1 the solution of the integral equation

$$g(x) = \frac{f_1(x) - f_0(x)}{f_1(x)} - \int K_1(x,y) g(y) \nu(dy). \tag{2.9}$$

We can put equation (2.9) into a symmetric form by making the substitution $g(x) = h(x)/\sqrt{f_1(x)}$. This yields the integral equation

$$h(x) = \frac{f_1(x) - f_0(x)}{\sqrt{f_1(x)}} - \int K_1^*(x,y)h(y)\nu(dy) \tag{2.10}$$

which has the symmetric kernel

$$K_1^*(x,y) = [f_1(x)f_1(y)]^{-\frac{1}{2}} \sum_{j=1}^{m} [f_1^j(x,y) + f_1^j(y,x)] - (2m+1)[f_1(x)f_1(y)]^{\frac{1}{2}}. \tag{2.11}$$

This form of the integral equation is useful since we may apply the Hilbert-Schmidt theory.

Consider now $\sigma_1^2(g_1)$, which we may write in the form

$$\begin{aligned}
\sigma_1^2(g_1) &= \int g_1^2(x)f_1(x)\nu(dx) + \sum_{j=1}^{m} \iint g_1(x)g_1(y)[f_1^j(x,y) + f_1^j(y,x)]\nu(dx)\nu(dy) \\
&\quad - (2m+1) \iint g_1(x)g_1(y)f_1(x)f_1(y)\nu(dx)\nu(dy) \\
&= \int g_1(x)f_1(x)\Bigg[g_1(x) + \int \Bigg\{ \frac{1}{f_1(x)} \sum_{j=1}^{m} [f_1^j(x,y) + f_1^j(y,x)] \\
&\quad - (2m+1)f_1(y) \Bigg\} g_1(y)\nu(dy) \Bigg] \nu(dx) \\
&= \int g_1(x)f_1(x)[g_1(x) + \int K_1(x,y)g_1(y)\nu(dy)]\nu(dx)
\end{aligned} \tag{2.12}$$

From this expression it can be seen that if g_1 solves the integral equation (2.7), then $\sigma_1^2(g_1) = \mu_1(g_1) - \mu_0(g_1)$, and thus $S_1(g_1) = \mu_1(g_1) - \mu_0(g_1)$ is the optimal value of S_1.

We have the following theorem.

Theorem 1. *If the process $\{X_i\}$ is m-dependent under H_1, then a necessary and sufficient condition for g_1 to maximize S_1 is that g_1 solve the integral equation (2.9). Furthermore, if g_1 solves (2.9) then $S_1(g_1) = \mu_1(g_1) - \mu_0(g_1)$.*

C. The solution of the integral equation for S_1

To apply the theory of Fredholm equations we require the following two conditions:

$$(a) \quad \int \frac{[f_1(x) - f_0(x)]^2}{f_1(x)} \nu(dx) < \infty$$

$$(b) \quad \iint |K_1^*(x,y)|^2 \nu(dx)\nu(dy) < \infty.$$

The condition (a) does not hold in many situations where the densities f_0, f_1 have tails to the left or right, as for example the Gaussian, lognormal, or Rayleigh densities. Therefore it may be necessary, in practice, to modify the tails of the densities in order to apply the theory. The condition that $f_0(x)/f_1(x)$ be bounded is sufficient but not necessary for condition (a) to hold. The condition (b) can be further characterized if the densities $f_1^j(x,y), j = 1, ..., m$ have the diagonal expansion [5]

$$f_1^j(x,y) = f_1(x)f_1(y) \sum_{n=0}^{\infty} a_n^{(j)} \theta_n(x)\theta_n(y). \tag{2.13}$$

where the functions $\{\theta_n\}$ are orthonormal in the sense that $\int \theta_m(x)\theta_n(x)f_1(x)\nu(dx) = \delta_{mn}$. Consider the terms in the expansion of $|K^*|^2$. We examine only the terms of the form $f_1^j(x,y)f_1^k(x,y)/[f_1(x)f_1(y)]$, the other terms being more obviously integrable. If we introduce the expansion (2.13) and apply the orthogonality relation, we have

$$\iint \frac{f_1^j(x,y)f_1^k(x,y)}{f_1(x)f_1(y)} \nu(dx)\nu(dy) = \sum_{n=0}^{\infty} a_n^{(j)} a_n^{(k)}. \tag{2.14}$$

Condition (b) will follow from the Schwarz inequality and (2.14)if $\sum |a_n^{(k)}|^2 < \infty$ for all k. Conditions (a) and (b) are sufficient to guarantee that the solution $h_1(x) = \sqrt{f_1(x)}g_1(x)$, if it exists, is square integrable, and this in turn implies that $\mathrm{E}_1 g_1^2(X_1) < \infty$ under our assumption that f_0/f_1 is bounded.

In the iid case, the kernel K_1^* reduces to $[f_1(x)f_1(y)]^{\frac{1}{2}}$, and it is easy to verify the solution

$$h(x) = \frac{cf_1(x) - f_0(x)}{\sqrt{f_1(x)}} \tag{2.15}$$

where c is an arbitrary constant. Note that the absolute term of the integral equation is of this form when $c = 1$. We may therefore define h_{iid} by

$$h_{\text{iid}}(x) = \frac{f_1(x) - f_0(x)}{\sqrt{f_1(x)}}$$

and write the integral equation as

$$h(x) = h_{\text{iid}}(x) + \int K_1^*(x, y)h(y)\nu(dy).$$

When the process is not iid, we can still obtain a series expansion of the solution to the integral equation using the Hilbert-Schmidt theory, provided we can find the eigenvalues and eigenvectors of the kernel. According to the theory, a unique solution exists provided the conditions (a) and (b) above hold and provided $+1$ is not an eigenvalue of the kernel. If 1 is an eigenvalue, a (non-unique) solution still exists, provided the absolute term h_{iid} is orthogonal to every eigenvector corresponding to the eigenvalue 1. We shall see that 1 is an eigenvalue of the kernel K_1^*, but that in most cases a solution still exists.

If we assume that the densities f_1^j, $j = 1, \ldots, m$ have the expansion (2.13) and we introduce this expansion into the kernel, we have

$$K_1^*(x, y) = \sqrt{f_1(x)f_1(y)}\left[(2m + 1) - 2\sum_{n=0}^{\infty}\left(\sum_{j=1}^{m} a_n^{(j)}\right)\theta_n(x)\theta_n(y)\right]. \tag{2.16}$$

Usually we will have $\theta_0(x) \equiv 1$ for such an expansion, and in such cases K_1^* will have eigenvalues $\{\lambda_n\}$ and eigenvectors $\{\phi_n\}$ given by

$$\lambda_0 = 1$$

$$\lambda_n = -2\sum_{j=1}^{m} a_n^{(j)} \qquad (n \geq 1)$$

$$\phi_n = \sqrt{f_1}\theta_n \qquad (n \geq 0).$$

Since $\lambda_0 = 1$, we must verify that h_{iid} is orthogonal to $\phi_0 = \sqrt{f_1}$, which is trivial:

$$\int h_{\text{iid}}(x)\phi_0(x)\nu(dx) = \int f_1(x) - f_0(x)\nu(dx) = 0.$$

If $\lambda_n \neq 1, n \geq 1$, we have the solution

$$h(x) = h_{\text{iid}}(x) + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\phi_n(x) + c\phi_0(x)$$

$$= h_{\text{iid}}(x) + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\theta_n(x)\sqrt{f_1(x)} + c\sqrt{f_1(x)}$$

with

$$c_n = \int h_{\text{iid}}(x)\phi_n(x)\nu(dx) = \int [f_1(x) - f_0(x)]\,\theta_n(x)\nu(dx)$$

and c an arbitrary constant. Therefore, the nonlinearity g (with $c = -1$) is given by

$$g(x) = -\frac{f_0(x)}{f_1(x)} + \sum_{n\geq 1} \frac{\lambda_n c_n}{1-\lambda_n}\theta_n(x). \tag{2.17}$$

D. Extension to ρ-mixing processes

The assumption of m-dependence in deriving the integral equation (2.9) is necessary to avoid problems with the interchanging of limits. In many situations, a strong mixing model is more appropriate. (See Appendix A for the terminology and notation used in relation to mixing processes). For example, Markov processes are not m-dependent. Realistically, if m is rather large, it is not possible to distinguish between a strong mixing process for which $\alpha_n > 0$ for all n and the m-dependent process which has mixing parameters $\tilde{\alpha}_n$ given by $\tilde{\alpha}_n = \alpha_n$ for $n \leq m$ and $\tilde{\alpha}_n = 0$ for $n > m$. Thus an m-dependent model might be substituted for a strong mixing model in many situations. However, it is of interest to investigate the possibility of generalizing the previous results by letting $m \to \infty$. Define $\sigma^2_{i,m}$ by

$$\sigma^2_{i,m}(g) = \text{Var}_i\, g(X_1) + 2\sum_{j=1}^{m} \text{Cov}_i[g(X_1), g(X_{j+1}]. \tag{2.18}$$

If $\{X_i\}$ is m_0-dependent under H_i, then $\sigma^2_{i,m} = \sigma^2_i$ for $m \geq m_0$. Define also

$$S_1^{(m)} = \frac{(\mu_1 - \mu_0)^2}{\sigma^2_{i,m}}. \tag{2.19}$$

The nonlinearity $g_1^{(m)}$ which solves the integral equation (2.9) maximizes the performance measure $S_1^{(m)}$. We would like to investigate the behavior of $g_1^{(m)}$ as $m \to \infty$. We shall assume that the processes are ρ-mixing with parameters satifying $\sum \rho_n < \infty$. The ρ-mixing condition is somewhat stronger than the strong mixing condition. A useful bound is $4\alpha_n \leq \rho_n$, which relates the parameters for the two conditions.

Actually, the results of [2] will apply here with trivial modification; however, our results here are easier to grasp intuitively, and generalize easier to the case of the other performance measures. The key to our result is the continuity of S_1 as a functional. Define the compact set $\mathcal{G}_1 = \{g : \mathrm{E}_1\, g(X_1) = 0, \mathrm{E}_1\, g^2(X_1) = 1\}$. Essentially, every nonconstant function g with finite second moment is represented by an element of $\mathcal{G}_1$ which is a scaled and shifted version of g. Since S_1 is invariant under shifting and scaling, we may characterize S_1 by considering its properties on $\mathcal{G}_1$. An important assumption we will have to make is that $\sigma_1^2(g) > 0$ for every $g \in \mathcal{G}_1$. Our result is first stated and then proved.

Theorem 2. *Suppose that the observation process $\{X_n\}$ is ρ-mixing under H_1 with $\sum \rho_n < \infty$, and that $\sigma_1^2(g) > 0$ for all $g \in \mathcal{G}_1$. Then there exists $g_1 \in \mathcal{G}_1$ such that $\sup_g S_1(g) = S_1(g_1)$. If $g_1^{(m)}$ solves the integral equation (2.9), then $S_1(g_1^{(m)}) \to S_1(g_1)$ as $m \to \infty$.*

Proof. From the definition of the mixing parameters ρ_n, it is a fact that $|\mathrm{Cov}_1[g(X_1), g(X_{j+1})]| < \rho_j$ for all $g \in \mathcal{G}_1$. Therefore the sum in (2.18) converges absolutely and uniformly as $m \to \infty$. It is easy to show that the functional $\sigma_{1,m}^2$ is continuous in the Hilbert space $\mathcal{L}_2(f_1)$, and the uniform convergence implies that σ_1^2 is continuous. Hence S_1 is a continuous function in $\mathcal{L}_2(f_1)$ whenever $\sigma_1^2 > 0$. The set $\mathcal{G}_1$ is compact and therefore there exists $g_1 \in \mathcal{G}_1$ such that $\sup_g S_1(g) = S_1(g_1)$. Furthermore, $S_1^{(m)}(g) \to S_1(g)$ uniformly for $g \in \mathcal{G}_1$ as $m \to \infty$. For $m \geq m_0$, $S_1^{(m)}$ is continuous, and therefore there exists in $\mathcal{G}_1$ a function $g_1^{(m)}$ such that $\sup_g S_1^{(m)}(g) = S_1^{(m)}(g_1^{(m)})$. We know

also that $g_1^{(m)}$ solves the integral equation (2.9). Now let $\epsilon > 0$. There exists an integer M such that for every $m \geq M$ and every $g \in \mathcal{G}_1$ we have $|S_1^{(m)}(g) - S_1(g)| < \epsilon$. Let $m \geq M$ be fixed. If $S_1^{(m)}(g_1^{(m)}) < S_1(g_1)$, then we must have $S_1^{(m)}(g_1) \leq S_1^{(m)}(g_1^{(m)}) < S_1(g_1)$. Otherwise, we have $S_1(g_1^{(m)}) \leq S_1(g_1) \leq S_1^{(m)}(g_1^{(m)})$. In either case, $|S_1^{(m)}(g_1^{(m)}) - S_1(g_1)| < \epsilon$. This implies that $|S_1(g_1^{(m)}) - S_1(g_1)| < 2\epsilon$. □

Our result here differs from that of [2] in that we work with properties of the performance measure itself, while Halverson and Wise work with the integral equations (2.9). If g were a nonconstant eigenvector of the kernel of the integral equation (2.9) corresponding to the eigenvalue $+1$, then from (2.12) we have $\sigma_1^2(g) = 0$. On the other hand, if $g \in \mathcal{G}_1$ and $\sigma_1^2(g) = 0$ then from the discussion above, we can reason that the optimization problem does not have a solution. Because the satisfaction of the integral equation (2.9) is both necessary and sufficient for a given nonlinearity to optimize the performance measure, it is clear that the condition on the eigenvalues of the kernel and the condition on σ_1^2 are equivalent.

The case of $\sigma_1^2(g) = 0$ is a degenerate case, and implies that $\frac{1}{n}\mathrm{Var}_1 T_n \to 0$ as $n \to \infty$, a situation which is more favorable for discrimination than otherwise. In this case the normalized test statistic converges to a constant under either hypothesis. Such a situation is quite unlikely to be encountered in practical problems, however.

III. The Performance Measures S_2 and S_3

The performance measure S_1 is conspicuously unequal in its treatment of the two hypotheses, resulting from the fact that the convergence of P_1 is optimized while the convergence of P_0 is neglected. In this section, we will derive performance measures for which the convergence of both error probabilities is optimized. Again we must assume that the distribution of the test statistic is approximately Gaussian for large sample sizes.

A. The derivation of the performance measure S_2

We now derive a performance measure S_2 which determines approximately the rate at which $P_1 \to 0$ with the threshold γ chosen so that $P_0 = P_1$. This is the asymptotic minimax problem: maximize the minimum of the convergence rates of P_0 and P_1.

The error probabilities for a single threshold test of the form (1.3) are given approximately by (2.1). Setting $P_0 = P_1$ we obtain $\gamma = (\mu_0\sigma_1 + \mu_1\sigma_0)/(\sigma_1 + \sigma_0)$. Then the common value of the error probabilities is

$$P_0 = P_1 = \Phi\left[-\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0 + \sigma_1}\right]. \tag{3.1}$$

Now define the performance measure S_2 by

$$S_2(g) = \frac{[\mu_1 - \mu_0]^2}{[\sigma_0(g) + \sigma_1(g)]^2}$$

From (3.1) we see that if the Gaussian approximation is good, then S_2 determines approximately the rate at which the error probabilities converge.

This performance measure has been derived in a slightly different way by using Chernoff bounds for Gaussian processes [3]. In that paper, it was assumed that the Chernoff bounds for the observation process $\{X_i\}$ can be approximated by the Chernoff bounds for Gaussian processes. Then S_2 can be derived by maximizing the minimum of the Chernoff exponents. Our derivation of the performance measure S_3 is a modification of this approach.

B. The optimal nonlinearity for S_2

We consider the optimization problem

$$\text{maximize} \quad S_2(g) = \frac{(\mu_1(g) - \mu_0(g))^2}{(\sigma_0(g) + \sigma_1(g))^2} \tag{3.2}$$

subject to the constraints that $\mathrm{E}_i g^2(X_1) < \infty$ for $i = 0, 1$. Sadowsky and Bucklew showed that the solution to the optimization problem is given by a certain nonlinear integral equation. The derivation is similar to the derivation of the integral equation (2.9) which gives the optimal nonlinearity

for S_1. If we assume m-dependence, then we may write the integral equation in the form

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)} - 2\lambda \int L(x,y)g(y)dy \tag{3.3}$$

where $\tau = (\sigma_1/\sigma_0)$ and the kernel L is given by

$$L(x,y) = \frac{(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)} \tag{3.4}$$

with

$$\tilde{K}_i(x,y) = \sum_{j=1}^{m} [f_i^j(x,y) + f_i^j(y,x)] - (2m+1)f_i(x)f_i(y). \tag{3.5}$$

Again we observe that λ determines the scaling of g. Thus the particular value of λ is not significant except that it must have the proper sign so that if g solves the integral equation (3.3), then $\mu_1(g) > \mu_0(g)$. Consider now

$$\begin{aligned} [\sigma_0(g) + \sigma_1(g)]^2 &= (1+\tau)\sigma_0^2(g) + (1+\tau^{-1})\sigma_1^2(g) \\ &= \int g(x)\Big\{ [(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)]g(x) \\ &\qquad \int [(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)]g(y)dy \Big\} dx \end{aligned} \tag{3.6}$$

If g solves the integral equation (3.3) for $\lambda = \frac{1}{2}$, then the expression in the braces in (3.6) reduces to $f_1(x) - f_0(x)$, and thus $[\sigma_0(g) + \sigma_1(g)]^2 = \mu_1(g) - \mu_0(g) > 0$. We shall therefore assign to λ the value $\frac{1}{2}$ and henceforth consider the integral equation

$$g(x) = \frac{f_1(x) - f_0(x)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)} - \int L(x,y)g(y)dy. \tag{3.7}$$

Furthermore, we observe that if g_2 solves the integral equation (3.7) then $S_2(g_2) = \mu_1(g_2) - \mu_0(g_2)$, a result which is similar to the one obtained in Section II for the optimal value of S_1.

We have the following theorem.

Theorem 3. *If the process $\{X_i\}$ is m-dependent under both H_0 and H_1, then a necessary and sufficient condition for g_2 to maximize S_2 is that g_2 solve the integral equation (3.7). Furthermore, if g_2 solves (3.7) then $S_2(g_2) = \mu_1(g_2) - \mu_0(g_2)$.*

In comparing the integral equation (3.7) with the integral equation (2.9), we note first of all that (3.7) is nonlinear because of the fact that τ is a function of g. Let us consider now what happens when τ varies. If τ is very small, then σ_1 is much smaller than σ_0, and thus the value of performance measure S_2 is very close to that of the performance measure S_0. In fact, $S_2 \to S_0$ as $\tau \to 0$. Now observe that the integral equation (3.7), when rescaled, converges as $\tau \to 0$ to the integral equation (2.9) which maximizes the performance measure S_1. This provides us with some insight to the relation between the performance measures S_0, S_1, and S_2 and the role that τ plays in the integral equation (3.7). We observe, for example, that there is a conflict of objectives for very small τ in that the value of the performance measure S_2 is approximately equal to that of S_0, while the integral equation (3.7) provides a nonlinearity which is close to the one which maximizes the performance measure S_1. A similar conflict occurs if τ approaches ∞, with the roles of S_0 and S_1 reversed. Of course, there is no conflict of objective if $S_0 \approx S_1$, but this implies that $\tau \approx 1$. Thus we expect that τ will have a "reasonable" value on the order of one. We find this to be the case in Section IV where a numerical solution to the integral equation (3.7) is found.

C. The solution of the integral equation for S_2

The equation (3.7) is nonlinear because τ is a function of g, and for this reason, finding a closed form solution is rather difficult. If, however, we had clairvoyance to know the correct value of τ, then we could find the solution g_2 by solving a linear integral equation. In fact, we might try to guess the value of τ, find the solution of the resulting linear integral equation, and then compute τ to verify if our guess was correct. This suggests an iterative method where the computed value of τ from the previous solution becomes the new value for τ at the next iteration of the procedure. This method was used to obtain a numerical solution to (3.7) for the results of Section V.

Although we cannot find a closed form solution to (3.7), we may treat τ as a constant whose value is unknown, and thereby extend the analysis relating to the equation (3.7). If we make the

substution

$$h(x) = g(x)\sqrt{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)},$$

we obtain the integral equation

$$h(x) = \frac{f_1(x) - f_0(x)}{\sqrt{w_\tau(x)}} + \int L^*(x,y)h(y)dy \tag{3.8}$$

where the symmetric kernel L^* is given by

$$L^*(x,y) = \frac{(1+\tau)\tilde{K}_0(x,y) + (1+\tau^{-1})\tilde{K}_1(x,y)}{\sqrt{w_\tau(x)w_\tau(y)}} \tag{3.9}$$

and where w_τ is defined by

$$w_\tau(t) = (1+\tau)f_0(t) + (1+\tau^{-1})f_1(t). \tag{3.10}$$

For a given value of τ, the integral equation (3.8) is a Fredholm equation of the second kind, provided we have the conditions

$$(a) \quad \int \frac{[f_1(x) - f_0(x)]^2}{w_\tau(x)} dx < \infty$$

$$(b) \quad \iint |L^*(x,y)|^2 dxdy < \infty.$$

These conditions imply that the solution h is square integrable, and then it follows that that $\mathrm{E}_i\, g^2(X_1) < \infty$ for $i = 0, 1$. Note that we do not require the condition that $f_0(x)/f_1(x)$ be bounded as we did for the integral equation (2.9). Condition (a) follows from the fact that $|f_1(x) - f_0(x)|/w_\tau(x)$ is bounded by $(1+\tau)^{-1} + (1+\tau^{-1})^{-1}$. To show that condition (b) holds, it suffices to show that

$$\iint \left| \frac{\tilde{K}_i(x,y)}{\sqrt{w_\tau(x)w_\tau(y)}} \right|^2 dxdy < \infty \tag{3.11}$$

since we may then apply the Minkowski inequality. If all the joint densities involved have the expansion (2.13), then the inequality (3.11) follows from a similar argument for the case of the kernel K_1^* in Section II. For example, consider the terms of the form

$$\frac{f_0^j(x,y)f_0^k(x,y)}{w_\tau(x)w_\tau(y)} \le \frac{f_0^j(x,y)f_0^k(x,y)}{(1+\tau)^2 f_0(x)f_0(y)}.$$

It was shown in Section II that such terms are integrable. Thus condition (b) holds. Since the integral equation (3.8) has a symmetric kernel, the Hilbert-Schmidt theory applies as in Section II. If the eigenvalues and eigenvectors of the kernel (3.9) are denoted by $\{\lambda_n\}$ and $\{\phi_n\}$, then a solution h_2 of the integral equation (3.8) has the expansion

$$\begin{aligned} h_2(x) &= h^*(x) + \sum_{n=0}^{\infty} \frac{\lambda_n c_n}{1-\lambda_n}\phi_n(x) \\ &= \sum_{n=0}^{\infty} \frac{c_n}{1-\lambda_n}\phi_n(x) \end{aligned} \tag{3.12}$$

where

$$h^*(x) = \frac{f_1(x) - f_0(x)}{\sqrt{w_\tau(x)}}$$

and

$$c_n = \int h^*(x)\phi_n(x)dx.$$

The solution is unique if and only if $\lambda_n \neq 1$ for all n. Since we do not have clairvoyance to know the true value of τ, the solution (3.12) is purely academic.

If the process is iid, then the kernel L from (3.4) has the simpler form

$$L(x,y) = -\frac{(1+\tau)f_0(x)f_0(y) + (1+\tau^{-1})f_0(x)f_1(y)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)}$$

and the integral equation (3.7) has the solution

$$g_{iid}(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{(1+\tau)f_0(x) + (1+\tau^{-1})f_1(x)} \tag{3.13}$$

where $B_0 = [(1+\tau)\mu_0 - 1]$ and $B_1 = [(1+\tau^{-1})\mu_1 + 1]$. There are three unknown quantities in the expression (3.13): τ, μ_0, and μ_1. These quantities can be found by solving the following system of nonlinear equations:

$$\begin{aligned} \mu_0 &= \int g_{iid}(x)f_0(x)dx \\ \mu_1 &= \int g_{iid}(x)f_1(x)dx \\ \tau &= \sqrt{\frac{\sigma_1^2(g_{iid})}{\sigma_0^2(g_{iid})}}. \end{aligned} \tag{3.14}$$

Note that $g_{iid}(x) + C$ solves (3.7) for an arbitrary constant C. We may therefore take an arbitrary value for either μ_0 or μ_1. In fact, with τ fixed the first two equations in (3.14) are linear in μ_0 and μ_1 and are singular. Therefore the system (3.14) does not have a unique solution.

D. The performance measure S_3

As we mentioned, the performance measure S_2 was derived in [3] by considering the Chernoff bounds for the error probabilities assuming Gaussian distributions. These bounds are the following:

$$\begin{aligned} P[T_n \geq n\gamma] &\leq \exp[-nI_i(\gamma)] \qquad \text{if } \mu_i < \gamma \\ P[T_n \leq n\gamma] &\leq \exp[-nI_i(\gamma)] \qquad \text{if } \mu_i > \gamma, \end{aligned} \tag{3.15}$$

where

$$I_i(\gamma) = \frac{(\mu_i - \gamma)^2}{2\sigma_i^2}. \tag{3.16}$$

The Chernoff bounds are asymptotically tight in the sense that

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n} \log P[T_n \geq n\gamma] &= I_i(\gamma) \qquad \text{if } \mu_i < \gamma \\ \lim_{n\to\infty} -\frac{1}{n} \log P[T_n \leq n\gamma] &= I_i(\gamma) \qquad \text{if } \mu_i > \gamma \end{aligned}$$

and hence they can provide good approximations to the error probabilites if n is large. Of course, if the distribution of T_n is only approximately normal, then the bounds given by (3.15) are only approximations of the true Chernoff bounds. Nevertheless, we shall proceed under the assumption that such approximations are acceptable. From (3.15) we see that if $n\gamma$ is the threshold and $\mu_0 < \gamma < \mu_1$, then I_i determines (approximately) the bound for the error probability P_i. Since a larger value for I_i results in a smaller bound for P_i, it is desirable to make both I_0 and I_1 as large as possible. It is obvious that I_i is a convex function of γ which takes its minimum value at $\gamma = \mu_i$. Thus as γ increases from μ_0 to μ_1, I_0 increases and I_1 decreases. Sadowsky and Bucklew proceeded from this point by maximizing $\min(I_0, I_1)$ and obtained the result that

$$S_2 = \max_{\mu_0 < \gamma < \mu_1} \min_{i=0,1} I_i(\gamma).$$

It is fairly straightforward, however, to show that

$$\max_{\mu_0<\gamma<\mu_1} [I_0(\gamma) + I_1(\gamma)] = \frac{[\mu_1 - \mu_0]^2}{[\sigma_0^2 + \sigma_1^2]}$$

and we therefore define the performance measure S_3 by

$$S_3 = \frac{[\mu_1 - \mu_0]^2}{[\sigma_0^2 + \sigma_1^2]}. \tag{3.17}$$

The optimization of S_3 leads to the linear integral equation

$$2\lambda g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + f_1(x)} - 2\lambda \int M(x,y)g(y)dy \tag{3.18}$$

where the kernel M is given by

$$M(x,y) = \frac{\tilde{K}_0(x,y) + \tilde{K}_1(x,y)}{f_0(x) + f_1(x)}.$$

It can easily be shown, by the same method as that used for the other performance measures, that if g_3 solves the integral equation (3.19), then $\sigma_0^2(g_3) + \sigma_1^2(g_3) = \mu_1(g_3) - \mu_0(g_3)$ so that the optimal value of S_3 is $S_3(g_3) = \mu_1(g_3) - \mu_0(g_3)$. Therefore, we will assign $\lambda = \frac{1}{2}$ and the integral equation becomes

$$g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + f_1(x)} - \int M(x,y)g(y)dy. \tag{3.19}$$

Therefore we have the following theorem.

Theorem 4. *If the process $\{X_i\}$ is m-dependent under both H_0 and H_1, then a necessary and sufficient condition for g_3 to maximize S_3 is that g_3 solve the integral equation (3.19). Furthermore, if g_3 solves (3.19) then $S_3(g_3) = \mu_1(g_3) - \mu_0(g_3)$.*

The integral equation (3.19) can be transformed into an integral equation with a symmetric kernel by making the substitution $h(x) = g(x)\sqrt{f_0(x) + f_1(x)}$ so that the Hilbert-Schmidt theory

applies as before. We shall not pursue this further. We shall, however, proceed to find the optimal nonlinearity for iid processes. If the processes are both iid, then the kernel M has the form

$$M(x,y) = -\frac{f_0(x)f_0(y) + f_1(x)f_1(y)}{f_0(x) + f_1(x)}$$

and the integral equation gives us immediately the form of the iid solution:

$$g_{iid}(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{f_0(x) + f_1(x)} \tag{3.20}$$

where $B_0 = \mu_0 - 1$ and $B_1 = \mu_1 + 1$. To find the unknown constants B_0, B_1, we subtitute for g_{iid} in the linear equations

$$\begin{aligned} \mu_0 = B_0 + 1 &= \int g_{iid}(x) f_0(x) dx \\ \mu_1 = B_1 - 1 &= \int g_{iid}(x) f_1(x) dx. \end{aligned} \tag{3.21}$$

The system (3.21) is in fact singular, so that we may take either B_0 or B_1 to be arbitrary. Therefore we shall arbitrarily take $B_0 = 0$ and this gives us the value

$$B_1 = \left[\int \frac{f_0(x) f_1(x)}{f_0(x) + f_1(x)} dx \right]^{-1}. \tag{3.22}$$

Thus the iid solution has been determined explicitly.

E. Extension to ρ-mixing processes

For ρ-mixing processes, we may prove results for S_2 and S_3 which are similar to those proved for S_1. In fact, the proof for S_3 is a simple modification for the proof for S_2; therefore, we shall include only the latter. Let $w = \frac{1}{2}(f_0 + f_1)$, let $\mathcal{G}_i = \{g : \mathrm{E}_i\, g(X_1) = 0, \mathrm{E}_i\, g^2(X_1) = 1\}$, and let $\mathcal{G} = \{g : \int g(x)w(x)\mu(dx) = 0, \int g^2(x)w(x)\mu(dx) = 1\}$. The result is stated in Theorem 5.

Theorem 5. *Suppose that the observation process $\{X_n\}$ is ρ-mixing under H_0 and H_1 with $\sum \rho_n < \infty$, and that $\sigma_i^2(g) > 0$ for all $g \in \mathcal{G}_i$, $i = 0, 1$. Then there exists $g_2 \in \mathcal{G}$ such that $\sup_g S_2(g) = S_2(g_2)$. If $g_2^{(m)}$ solves the integral equation (3.7), then $S_2(g_2^{(m)}) \to S_2(g_2)$ as $m \to \infty$.*

Proof. The ρ-mixing condition implies that $\sigma^2_{i,m}(g)$ converges uniformly for $g \in \mathcal{G}_i$ and thus σ^2_i is continuous in the Hilbert space $\mathcal{L}_2(f_i)$. If $g \in \mathcal{G}$ then for either $i = 0$ or $i = 1$ there exist $a > 0$ and b such that $ag + b \in \mathcal{G}_i$. Therefore for all $g \in \mathcal{G}$, the denominator in S_2 is positive. Because convergence in $\mathcal{L}_2(w)$ implies convergence in $\mathcal{L}_2(f_i)$, we can deduce that μ_0, σ_0^2, μ_1, and σ_1^2 are continuous in $\mathcal{L}_2(w)$, and hence S_2 is continuous in $\mathcal{L}_2(w)$ whenever the denominator is positive. The set $\mathcal{G}$ is compact in $\mathcal{L}_2(w)$, so there exists $g_2 \in \mathcal{G}$ such that $\sup_g S_2(g) = S_2(g_2)$.

Now define

$$S_2^{(m)} = \frac{(\mu_1 - \mu_0)^2}{(\sigma_{0,m} + \sigma_{1,m})^2}. \tag{3.23}$$

The argument given in the proof of Theorem 2 applies to prove that $S_2(g_2^{(m)}) \to S_2(g_2)$ if we can show that $S_2^{(m)}(g) \to S_2(g)$ uniformly for $g \in \mathcal{G}$ as $m \to \infty$. Let $g \in \mathcal{G}$ and let the constants $a_i > 0$ and b_i be such that $g_i = a_i g + b_i \in \mathcal{G}_i$. If such constants do not exist, then $a_i > 0$ and b_i may be arbitrary. Then we can write

$$\left|S_2(g) - S_2^{(m)}(g)\right| = \left|\left[\frac{\mu_1(g_1) - \mu_0(g_1)}{c\sigma_0(g_0) + \sigma_1(g_1)}\right]^2 - \left[\frac{\mu_1(g_1) - \mu_0(g_1)}{c[\sigma_0(g_0) + \epsilon_0] + [\sigma_1(g_1) + \epsilon_1]}\right]^2\right| \tag{3.24}$$

where $c = a_1/a_0$, and where ϵ_0 and ϵ_1 converge to 0 uniformly as $m \to \infty$. For fixed values of ϵ_0 and ϵ_1, the right-hand side of (3.24) attains its maximum as a function of c, and this maximum converges to 0 as ϵ_0 and ϵ_1 converge to 0. Hence convergence of $S_2^{(m)}$ is uniform on $\mathcal{G}$. This completes the proof. □

IV. Minimax Robustness

A. Preliminaries

In this section, we obtain results in robustness for the performance measures S_0, S_1, and S_3. An attempt to obtain results for S_2 leads to intractable expressions due to the nonlinear form of the corresponding integral equation.

The problem of minimax robustness is one of finding the least favorable distributions F_0^* and F_1^* in given uncertainty classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$. These least favorable distributions together with the robust nonlinearity g^* satisfy the inequalities

$$S(g, F_0^*, F_1^*) \leq S(g^*, F_0^*, F_1^*) \leq S(g^*, F_0, F_1) \tag{4.1}$$

where $F_0 \in \mathcal{Q}_0$, $F_1 \in \mathcal{Q}_1$, g is an arbitrary allowable (second order) nonlinearity, and S is a generic performance measure. Throughout this section, the performance measures are written as functions of the nonlinearity and the distributions, as in (4.1). We shall also use the notation $\mu(g; f) = \int g(x)f(x)dx$ and $\sigma^2(g; f) = \int g^2(x)f(x)dx - \mu^2(g; f)$, which indicates clearly the dependence of μ and σ on the marginal density f.

We define $\mathcal{Q}_0$ and $\mathcal{Q}_1$ by restrictions on the marginal and bivariate joint densities, as is appropriate for our performance measures. Let $\tilde{f}_0$ and $\tilde{f}_1$ denote the nominal marginal densities under H_0 and H_1, respectively. We require that for every distribution $F_i \in \mathcal{Q}_i$, the marginal density be of the form $f_i = (1 - \epsilon_i)\tilde{f}_i + \epsilon_i h$. This restriction on the marginal densities defines ϵ-contamination classes. Other classes of marginal densities may also be considered, such as the total-variation classes, the bounded classes, and the p-point classes. (See Lemma 7 and the comment following it).

The nominal distribution $\tilde{F}_i$ for the class $\mathcal{Q}_i$ is assumed to be i.i.d. Other distributions in $\mathcal{Q}_i$, in addition to satisfying the restriction on the marignal densities, must also satisfy the following restriction on the bivariate joint densities.

$$\sup_g \frac{|\mathrm{Cov}[g(X_1), g(X_{j+1})]|}{\sqrt{\mathrm{Var}\, g(X_1)\, \mathrm{Var}\, g(X_{j+1})}} \leq r_j \tag{4.2}$$

where g ranges over all measurable functions satisfying $\mathrm{E}\, g^2(X_1) < \infty$. Since we assume stationarity, the denominator is $\mathrm{Var}\, g(X_1)$. The r parameters are dominated by the ρ parameters which define the ρ-mixing condition; therefore, ρ-mixing processes will be included in $\mathcal{Q}_i$. It will turn out that the

least favorable distributions are in fact ρ-mixing with the ρ parameters equal to the r parameters, so that there is no need for concern if there are process distributions in the uncertainty classes defined here which are not ρ-mixing.

In applications, one may wish to consider a model which involves ρ-mixing processes, such as the processes which are assumed in the hypotheses of Theorems 2 and 5. If the r parameters for the bounds (4.2) are chosen to be equal to or larger than the ρ parameters, and the nominal marginal densities are taken to be the marginal densities of the models, then the model densities will be contained in the uncertainty classes. Often it is difficult to determine the ρ parameters or the r parameters from given densities. If the bivariate densities of the distributions have the diagonal expansion (2.13) with the orthonormal functions $\{\theta_i\}$ being polynomials, then the following result is useful for determining the r parameters.

Proposition 6. *Suppose the bivariate density $f^{(2)}$ has a diagonal expansion $f^{(2)}(x,y) = f(x)f(y)\sum_{n=0}^{\infty} a_n\theta_n(x)\theta_n(y)$ where $\{\theta_n\}$ form a complete orthonormal set of polynomials with the weight function f. Then*

$$\sup_g \frac{|\mathrm{Cov}[g(X),g(Y)]|}{\mathrm{Var}\, g(X)} = \max(a_1,a_2) \tag{4.3}$$

where the supremum is taken over all measurable g such that $\mathrm{E}\, g^2(X) < \infty$.

Proof. Since $\int g^2(x)f(x)dx < \infty$, g has a Fourier series $g(x) = \sum_{n=0}^{\infty} b_n\theta_n(x)$. It is easy to verify that

$$\frac{|\mathrm{Cov}[g(X),g(Y)]|}{\mathrm{Var}\, g(X)} = \frac{\left|\sum_{n\geq 1} b_n^2 a_n^{(j)}\right|}{\sum_{n\geq 1} b_n^2} \tag{4.4}$$

The supremum of the left side of (4.4) is obtained by $g = \theta_i$ where i is such that $|a_i| = \max\{|a_n|, n \geq 1\}$. If the orthonormal functions $\{\theta_n\}$ are polynomials then this maximum coefficient occurs as either $a_1^{(j)}$ or $a_2^{(j)}$. To show this, we require a fact from [12] that for any such diagonal expansion

in which the orthonormal functions are polynomials, there exists a probability density function h_j having support in the interval $[-1,1]$ such that $a_n = \int_{-1}^{1} t^n h(t)dt$. Then for $n > 2$ it is obvious that

$$|a_n| \leq \int_{-1}^{1} |t|^n h(t)dt \leq \int_{-1}^{1} t^2 h(t)dt = |a_2|.$$

so that the assertion holds. □

As a consequence of Proposition 6, the condition which defines the uncertainty class becomes a condition on the process $\{X_i\}$ directly, rather than a condition on tranformed processes $\{g(X_i)\}$ as in (4.2). Two important classes of densities which have diagonal expansions of the form required in Proposition 6 are the Gaussian and gamma densities. Note that Proposition 6 can be applied in an indirect way to processes which are memoryless transformations of processes which have a diagonal expansion (memoryless transformations of Gaussian processes, for example). The calculation of a_1 requires knowledge of the first and second order moments (a_1 is, in fact, the correlation coefficient) while the calculation of a_2 requires knowledge of the moments up to order four.

The condition (4.2) is from [9], and we shall adapt some of the results therein. For a given marginal distribution function F, equality holds in (4.2) for all g if the bivariate distribution function is

$$F^j(x,y) = (1-r_j)F(x)F(y) + r_j F(x \wedge y), \quad j = 1,2,\ldots \tag{4.5a}$$

where $x \wedge y$ is the minimum of x and y. If the distribution function F has a density f, then we may write for the bivariate densities

$$f^j(x,y) = (1-r_j)f(x)f(y) + r_j\delta(x-y)f(x), \quad j = 1,2,\ldots. \tag{4.5b}$$

In the work that follows, we will show that F_0^* and F_1^* have bivariate distributions of the form (4.5). The robustness problem therefore reduces to one which involves only the marginal distributions.

That there exist processes which have bivariate distributions given by (4.5) is shown by two constructions in [9]. Let $\{\theta_n\}$ be a sequence of nonnegative real numbers such that $\sum_{i=1}^{\infty} \theta_i = 1$. Sadowsky's constructions allow r-sequences of the form $r_j = \sum_{m=0}^{\infty} \theta_m \theta_{j+m}$ or $r_j = \sum_{m=j+1}^{\infty} \frac{m-j}{m} \theta_m$. Note that if the condition (4.2) is satisfied, then $\sigma_i^2(g) = (1 + 2R_i)\mathrm{Var}_i\, g(X_1)$ where $R_i = \sum r_j$. For this reason, the sum R_i is relevant for our work, while the particular sequence $\{r_j\}$ is not.

B. The least favorable processes

We have given the least favorable bivariate densities in (4.5). Now we show that these densities are, in fact, least favorable. Suppose f_0^* and f_1^* are least favorable univariate densities for the performance measure $\hat{S}$:

$$\hat{S}(g, f_0^*, f_1^*) \le \hat{S}(g^*, f_0^*, f_1^*) \le \hat{S}(g^*, f_0, f_1). \tag{4.6}$$

In this case, $\hat{S}$ could be $[\mu_1(g; f_1) - \mu(g; f_0)]^2/\sigma^2(g; f_1)$, which represents a univariate version of S_1, or it could be $[\mu(g; f_1) - \mu(g; f_0)]^2/[(1+2R_0\sigma^2(g; f_0) + (1+2R_1)\sigma^2(g; f_1)]$ which corresponds to S_3. We shall assume the latter, though it is simple to modify the argument to accomodate the former. We now show that the distributions given by (4.5) with the marginal densities f_0^* and f_1^* are least favorable by showing that (4.6) implies (4.1). The key to the proof is the fact that equality holds in (4.2) for every g when the bivariate densities are given by (4.5). Let F_i^* denote any process distribution which has marginal density f_i^* and bivariate densities (4.5). Let f_i denote an arbitrary marginal density from $\mathcal{Q}_i$, let $\hat{F}_i$ have marginal density f_i and bivariate densities (4.5), and let F_i be any distribution in $\mathcal{Q}_i$ with marginal f_i. Then we have from (4.6)

$$S_3(g, F_0^*, F_1^*) = \hat{S}(g, f_0^*, f_1^*) \le \hat{S}(g^*, f_0^*, f_1^*) = S_3(g^*, F_0^*, F_1^*)$$

which is the left inequality in (4.1). Furthermore,

$$S_3(g^*, F_0^*, F_1^*) = \hat{S}(g^*, f_0^*, f_1^*) \le \hat{S}(g^*, f_0, f_1) = S_3(g^*, \hat{F}_0, \hat{F}_1) \le S_3(g^*, F_0, F_1)$$

which is the right inequality.

The following marginal densities, defined in terms of the nominal densities $\tilde{f}_0$ and $\tilde{f}_1$, will be called the *Huber-Strassen least favorable densities*:

$$\begin{aligned} p_0(x) &= \begin{cases} (1-\epsilon_0)\tilde{f}_0(x) & \text{if} \quad \tilde{f}_1(x)/\tilde{f}_0(x) < c'' \\ (1/c'')(1-\epsilon_0)\tilde{f}_1(x) & \text{if} \quad \tilde{f}_1(x)/\tilde{f}_0(x) \geq c'' \end{cases} \\ p_1(x) &= \begin{cases} (1-\epsilon_1)\tilde{f}_1(x) & \text{if} \quad \tilde{f}_1(x)/\tilde{f}_0(x) > c' \\ c'(1-\epsilon_1)\tilde{f}_0(x) & \text{if} \quad \tilde{f}_1(x)/\tilde{f}_0(x) \leq c' \end{cases} \end{aligned} \tag{4.7}$$

where the constants c' and c'' are chosen such that the functions are valid probability densities (i.e. they integrate to 1). The Huber-Strassen densities have appeared frequently as the solution to various minimax robustness problems. Lemma 7 is the basis for many such applications.

Lemma 7. For $i = 0, 1$, let $\mathcal{P}_i$ be the class of all probability density functions of the form $f = (1-\epsilon_i)\tilde{f}_i + \epsilon_i h$, where $\tilde{f}_i$ is fixed and h is arbitrary, and let Ψ be any convex function. If p_0 and p_1 are the Huber-Strassen least favorable densities corresponding to $\tilde{f}_0$ and $\tilde{f}_1$, then the inequality

$$\int \Psi\left[\frac{p_1(x)}{p_0(x)}\right] p_0(x)dx \leq \int \Psi\left[\frac{f_1(x)}{f_0(x)}\right] f_0(x)dx$$

holds for all marginal densities $f_0 \in \mathcal{P}_0$ and $f_1 \in \mathcal{P}_1$.

Proof. It has been shown in [7] that the least favorable densities in terms of risk for the classes $\mathcal{P}_0$ and $\mathcal{P}_1$ are the Huber-Strassen densities. The proof then follows as a corollary to Lemma 1 in [15]. □

In addition to the ϵ-contamination classes, the Huber-Strassen densities are also least favorable in terms of risk for at least three other uncertainty classes: the total variation classes [7], bounded classes [17], and p-point classes [18]. Thus Lemma 7 holds as well if the classes $\mathcal{P}_0$, $\mathcal{P}_1$ are of one of these types.

The main result of this section is stated in the following theorem.

Theorem 8. *For the performance measures S_0, S_1, and S_3, the least favorable process distributions F_0^*, F_1^* in the classes $\mathcal{Q}_0$, $\mathcal{Q}_1$ are such that their marginal densities are the Huber-Strassen densities (4.7) and their bivariate joint distributions are defined by (4.5).*

Proof. See Appendix C.

V. Numerical Results

The nonlinearities which are given as solutions to the integral equations in the preceding sections are optimal in the sense that they optimize the various performance measures. We justified the performance measures by showing that under certain conditions, including large sample sizes, they imply small error probabilities. This final section will present approximations to the error probabilities which were generated by computer simulations for several examples, and thereby provide further justification of the use of various nonlinearities we have derived.

A. Descriptions of the examples

The particular marginal densities which we shall assume throughout are Rayleigh and log-normal, given by

$$f_0(x) = \frac{x}{\theta} \exp\left\{-\frac{x^2}{2\theta}\right\} \tag{5.1}$$

$$f_1(x) = \frac{1}{x\sqrt{2\pi\lambda_2}} \exp\left\{-\frac{(\log x - \lambda_1)^2}{2\lambda_2}\right\}, \tag{5.2}$$

respectively. For our simulations, we have taken $\theta = 4$, $\lambda_1 = 0.8$, and $\lambda_2 = 0.25$. These values were chosen so that $\mathrm{E}_0 X_1 = \mathrm{E}_1 X_1$ and $\mathrm{E}_0 X_1^2 = \mathrm{E}_1 X_1^2$. The bivariate densities are

$$f_0^j(x,y) = \frac{xy}{(1-\rho_j^2)\theta^2} \exp\left\{-\frac{x^2+y^2}{2(1-\rho_j^2)\theta}\right\} I_0\left\{\frac{\rho_j xy}{(1-\rho_j^2)\theta}\right\} \tag{5.3}$$

$$f_1^j(x,y) = \left[2\pi xy\lambda_2(1-\rho_j^2)^{\frac{1}{2}}\right]^{-1} \times \exp\left\{-\frac{(\log x - \lambda_1)^2 - 2\rho_j(\log x - \lambda_1)(\log y - \lambda_1) + (\log y - \lambda_1)^2}{2\lambda_2(1-\rho_j^2)}\right\} \tag{5.4}$$

with I_0 being the modified Bessel function of order 0. In general, the n-dimensional Rayleigh density requires an $(n-1)$-fold integration, and thus the LRT for Rayleigh vs. lognormal processes lacks a closed form expression.

The parameters ρ_j which are in both expressions (5.3) and (5.4) are actually the correlation coefficients of the underlying Gaussian processes from which the Rayleigh and lognormal processes are derived. In either case, the densities factor into a product of marginal densities when $\rho_j = 0$. We refer to the ρ_j's as the correlation parameters. For our examples, the correlation parameters will be given by decaying exponential sequences with time constant τ_i under H_i. For example, under H_0 the bivariate densities are given by (5.3) with $\rho_j = e^{(-j/\tau_0)}$. It can be shown that the processes we have specified are Markov processes. In the particular examples we present, the correlation time constants will take different values to illustrate the change in performance which results from a change in the strength of the correlation.

B. The calculation of the nonlinearities

For each example, we evaluate the performance of five different nonlinearities g_i, $i = 0, \ldots, 4$. The nonlinearity $g_4 = \log(f_1/f_0)$ is the optimal iid nonlinearity derived from the iid LRT. The other nonlinearities $g_0, \ldots, g_3$ are optimal under the respective performance measures $S_0, \ldots, S_3$, and are determined by the solutions of the integral equations of the preceding sections.

In obtaining (numerical) solutions to the integral equations, one must take care in the selection of two sets of parameters. First, one must decide on the number m of terms in the sums in the kernels. For a kernel such as (3.4), one may wish to assign separate values m_0 and m_1 for the two sums. Our method for determining these values was to set $\rho_{\min} = 0.1$ and then let m_i be the smallest value such that $\rho_{m_i} < \rho_{\min}$. We found that the same results were obtained when $\rho_{\min} = 0.01$, which corroborates Theorems 2 and 5.

The second set of parameters is $\{x_{\min}, x_{\max}\}$ which define the integration region $[x_{\min}, x_{\max}]$.

In situations where the observations can take on only bounded values, the selection of these parameters is trivial. If the marginal densities have unbounded support, however, such as the Rayleigh and lognormal densities, a finite region must be chosen in order to obtain a numerical solution to the integral equation. For our example, the integral equations corresponding to S_0 and S_1 will not have solutions unless the tails of the densities are modified. We selected $x_{\min}$ and $x_{\max}$ so that $P\{X_1 < x_{\min}\} < \epsilon$ and $P\{X_1 > x_{\max}\} < \epsilon$ under either hypothesis, with $\epsilon = 5 \times 10^{-5}$. This yields $x_{\min} = 0.02$ and $x_{\max} = 15.7$. Thus we have neglected the tail regions of extremely small total probability.

Figures 1 and 2 show the graphs of the nonlinearities which were computed as numerical solutions of the integral equations when the time constants had the values $\tau_0 = 13.029$ and $\tau_1 = 130.29$. The magnitudes of g_0, g_1, and g_4 become very large near the endpoints of the integration region. This is shown more clearly in Figures 3 and 4 which contain logarithmic plots. As could be expected from observing the form of g_0, $\sigma_0^2(g_0)$ is very small while $\sigma_1^2(g_0)$ is very large. The opposite situation is true for g_1, while for g_4 both $\sigma_0^2(g_4)$ and $\sigma_1^2(g_4)$ are quite large. Since g_2 does not take large values anywhere in the interval, $\sigma_0^2(g_2)$ and $\sigma_1^2(g_2)$ are both relatively moderate and of the same order of magnitude. The nonlinearity g_3 is similar to g_2 is this context.

C. Simulation results

Figures 5–8 show the ROCs for each of the nonlinearities for four different examples. The parameters for the marginal densities are the same for each of the examples, the time constants of the correlation parameters being the only varying parameters. The sample size for each trial was $n = 1000$, and 10,000 trials were performed to generate each ROC.

For the example of Figure 5, the ordering of the nonlinearities from best to worst at the point of the ROC where the error probabilities are equal is g_4, g_2, g_3, g_1, g_0. The situation here is one of weak correlation under both hypotheses, the time constants being $\tau_0 = 13.029$ and $\tau_1 = 13.029$.

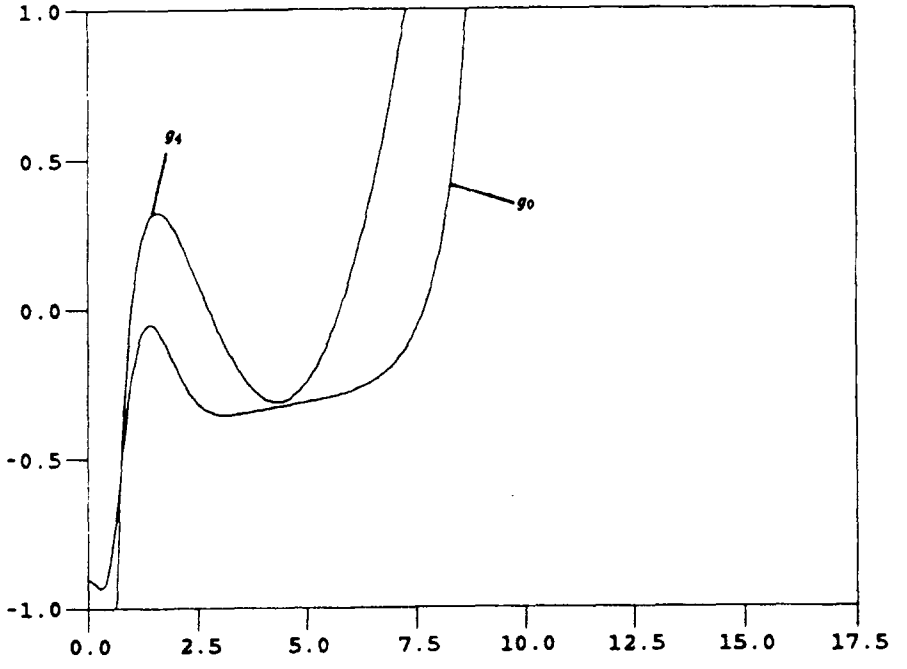

Figure 1. Linear graphs of the nonlinearities g_0 and g_1 when $\tau_0 = 13.029$ and $\tau_1 = 130.29$.

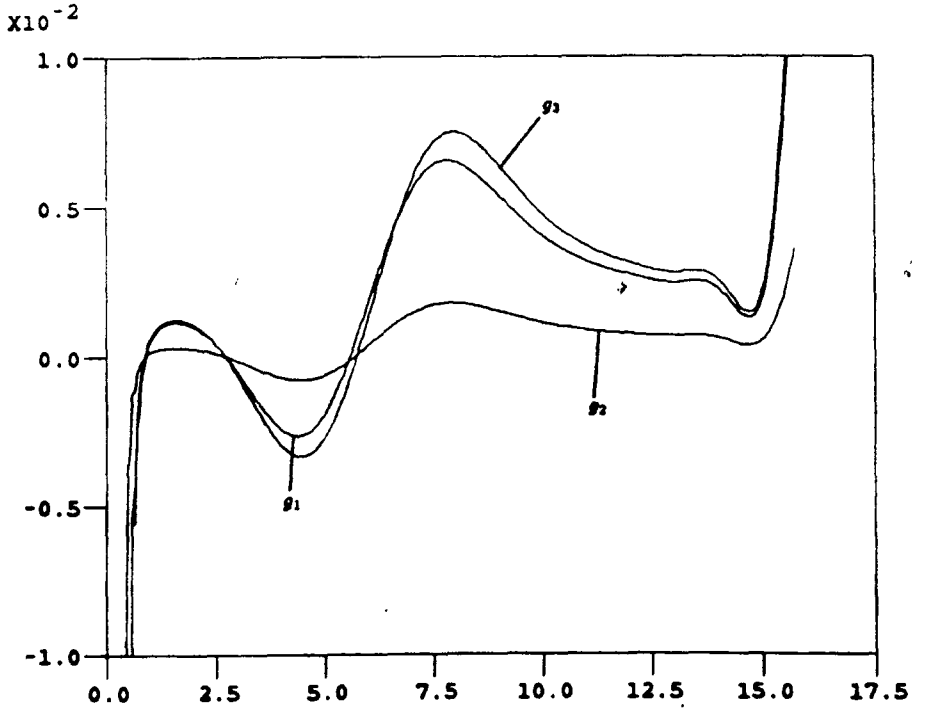

Figure 2. Linear graphs of the nonlinearities g_1, g_2, and g_3 when $\tau_0 = 13.029$ and $\tau_1 = 130.29$.

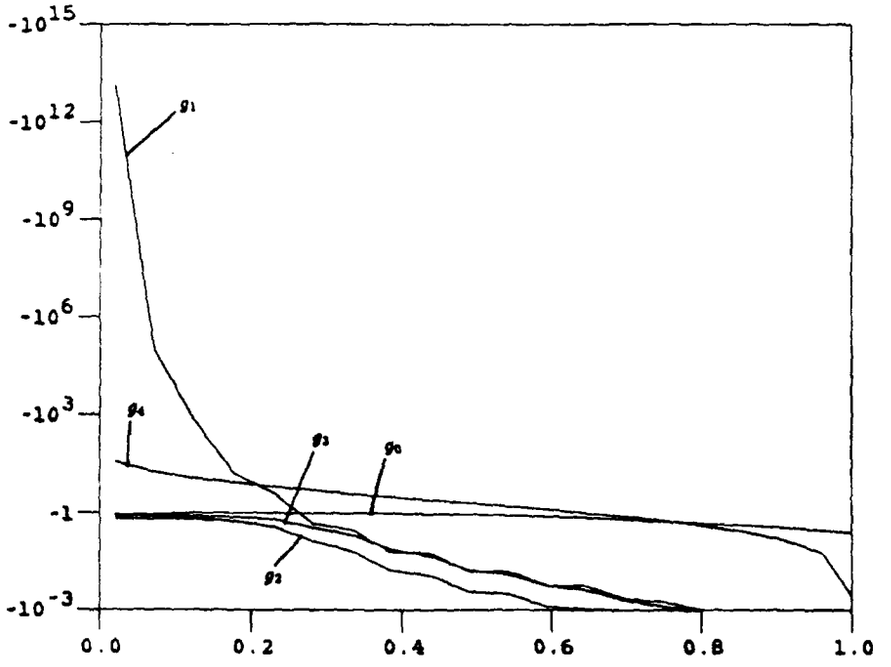

Figure 3. Logarithmic graphs of the left tails of the nonlinearities plotted in Figures 1 and 2.

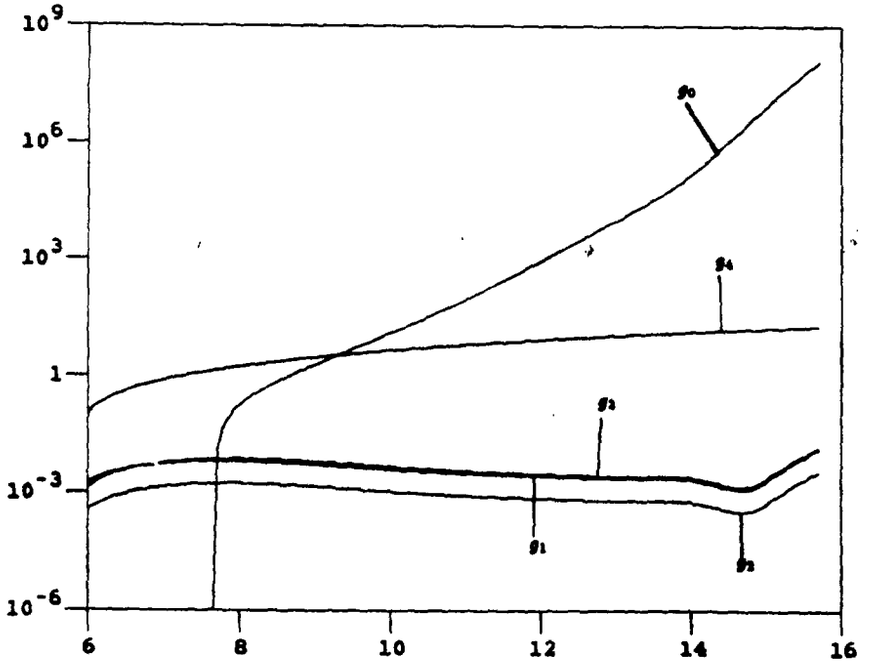

Figure 4. Logarithmic graphs of the right tails of the nonlinearities plotted in Figures 1 and 2.

That the optimal iid nonlinearity performs best is therefore not alarming.

For the next example shown in Figure 6, the situation is such that the correlation is strong under H_1 and weak under H_0, the time constants being $\tau_0 = 13.029$ and $\tau_1 = 130.29$. The ordering here is g_1, g_2, g_3, g_4, g_0. We see a significant improvement over the optimal iid nonlinearity.

Assigning the values $\tau_0 = 130.29$ and $\tau_1 = 13.029$, we obtain the ROCs shown in Figure 7. Although the value of τ_0 is significantly larger than its value in the first example, the correlation of the Rayleigh process is still not significantly strong. Comparing the results here with those of the first example (Figure 5), we see that the ordering of the nonlinearities at the point where the error probabilities are equal is unchanged; however the performance of each of the nonlinearities is degraded as a result of the stronger correlation.

In the final example shown in Figure 8, the time constants have values $\tau_0 = 130.29$ and $\tau_1 = 130.29$. The situation of the ROCs is similar to that of the second example (Figure 6), in that the ordering is the same. The performance in this example is degraded from that of the second example as a result of the increase in the correlation under H_0.

From the results we have examined so far, we might inquire as to the reason why g_0 performs so poorly. To explain this, we must first consider the fact that the correlation which is observed in the Rayleigh process is much weaker than the correlation observed in the Gaussian processes which generate them. (Recall that $Z = \sqrt{X^2 + Y^2}$ is Rayleigh when X and Y are iid zero-mean Gaussian). Therefore, even with the time constants as in Figure 7, the correlation of the Rayleigh process is not significantly strong. Second, recall the formulation under which g_0 has been derived, that of minimizing the rate of convergence of P_0 when P_1 is constrained. Therefore, we should consider the region of the ROC where P_0 is relatively small and P_1 relatively large. In Figure 7, we see that g_0 performs best in this region. This same phenomenon is observed even more clearly in g_1. In Figures 6 and 8, we see that g_1 performs significantly better than the others in the region where P_1 is small.

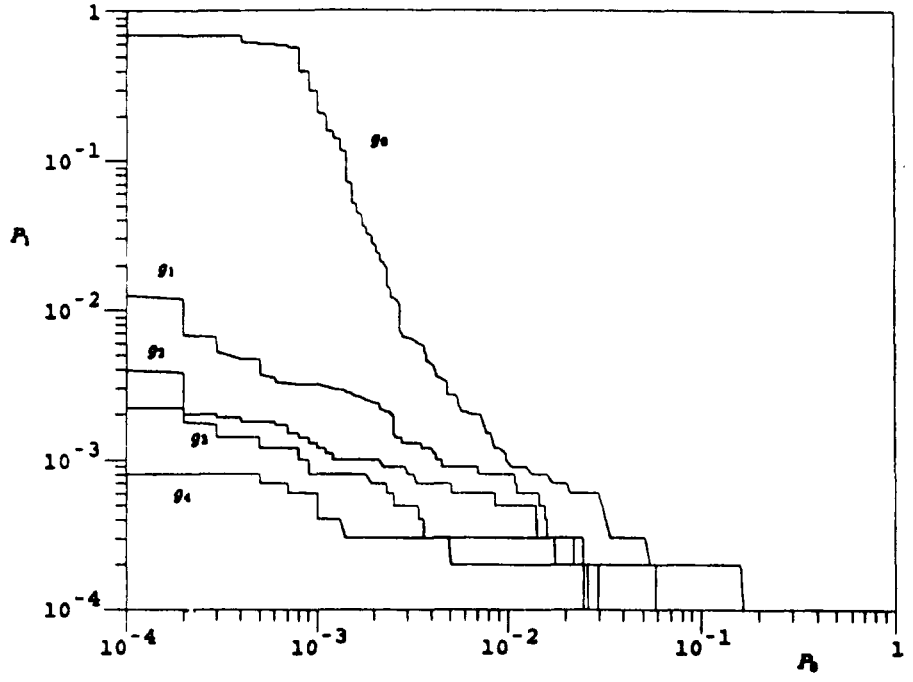

Figure 5. Receiver operating characteristics (ROCs) for $\tau_0 = 13.029$ and $\tau_1 = 13.029$ with $n = 1000$ samples.

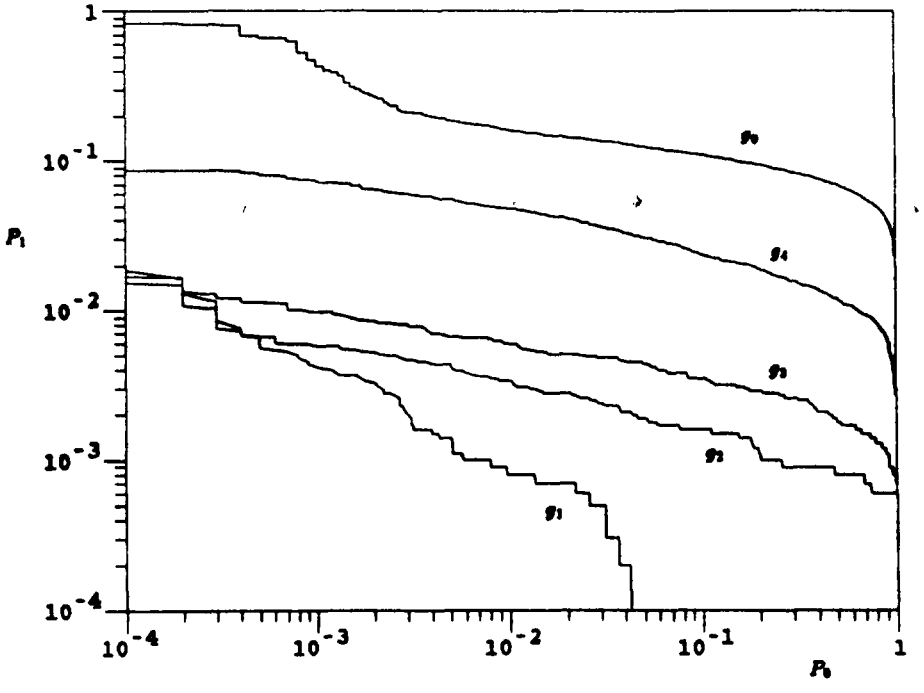

Figure 6. ROCs for $\tau_0 = 13.029$ and $\tau_1 = 130.29$ with $n = 1000$.

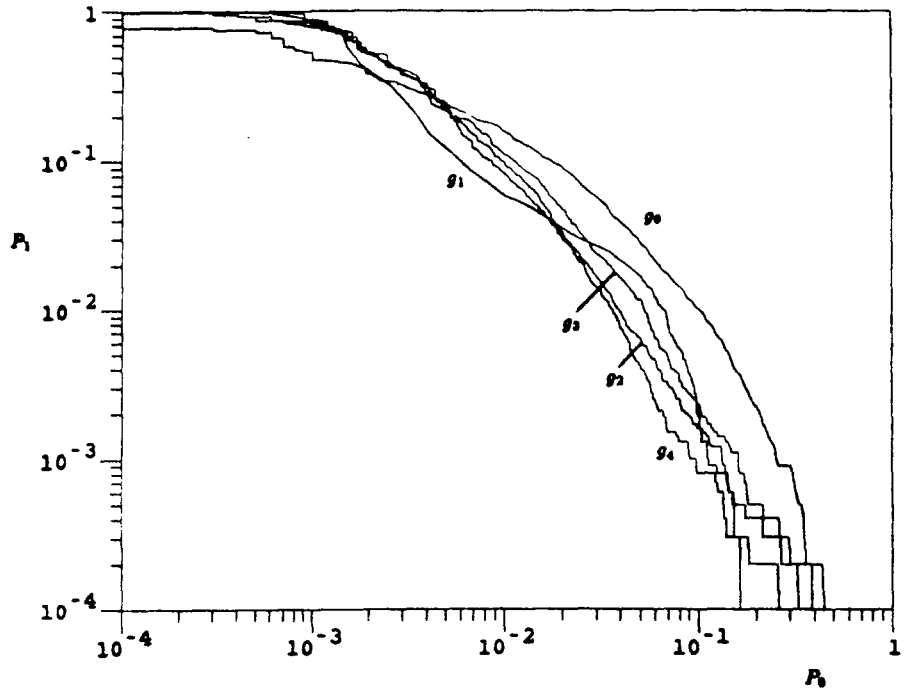

Figure 7. ROCs for $\tau_0 = 130.29$ and $\tau_1 = 13.029$ with $n = 1000$.

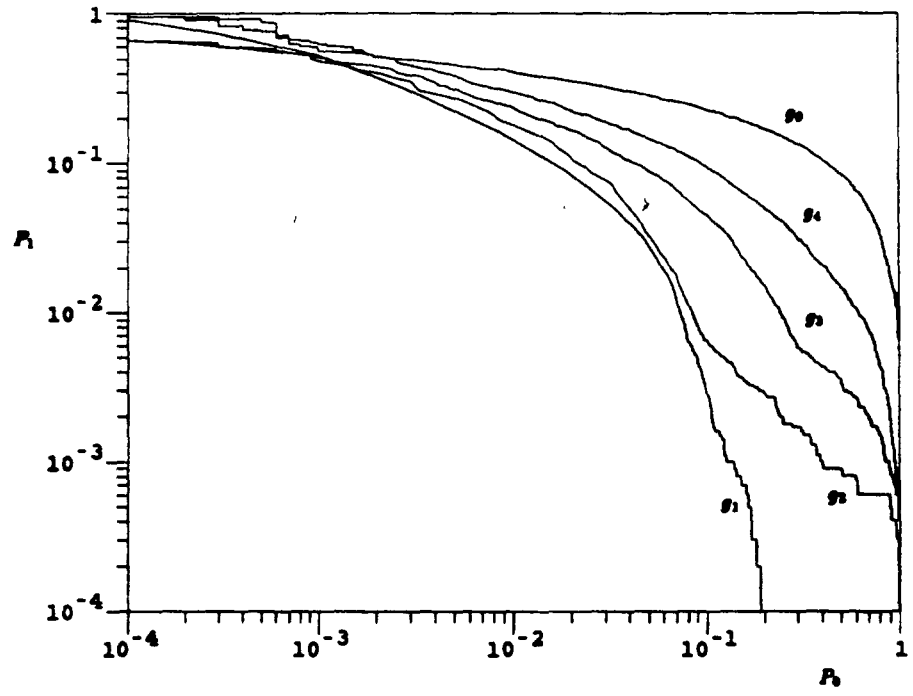

Figure 8. ROCs for $\tau_0 = 130.29$ and $\tau_1 = 130.29$ with $n = 1000$.

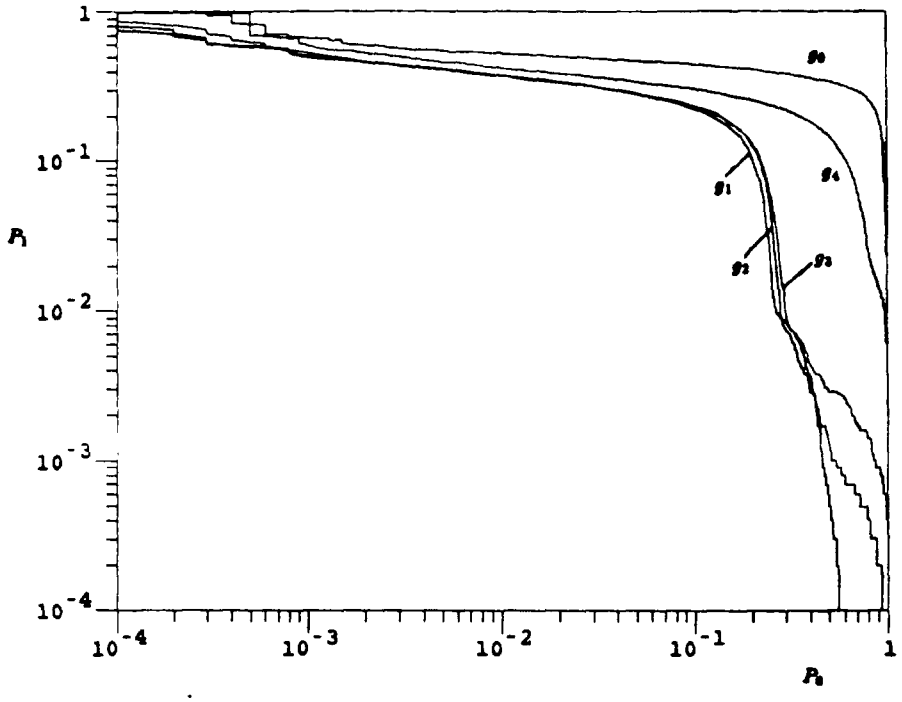

Figure 9. ROCs for $\tau_0 = 13.029$ and $\tau_1 = 130.29$ with $n = 100$.

To ascertain the effects of a smaller sample size, we include also Figure 9, in which the sample size n has the value 100. The ordering of the nonlinearities remains essentially unchanged, although the performance is degraded as a result of the decrease in the sample size.

VI. Conclusion

Four different performance measures have been presented in this paper, each giving rise to a different nonlinearity. A fifth possible nonlinearity is the optimal iid nonlinearity. From our numerical results in Section V, we might conclude that it is difficult to predict which of the nonlinearities will perform best in a given situation. Based on our numerical results, however, we might suggest the following heuristics. In a situation of weak correlation, the optimal iid nonlinearity will most likely perform best, owing to the fact that the other nonlinearities are designed to be optimal with respect to the various performance measures and not the error probabilities directly. In cases of

strong correlation under one of the hypotheses, say H_1, and weak correlation under H_0, according to our experience the nonlinearity which maximizes S_1 is likely to perform best. This is especially true if one is concerned about the region of the ROC where P_1 is small relative to P_0. The nonlinearities which maximize the performance measures S_2 and S_3 have performed consistently in the middle relative to the other nonlinearities which we have considered. In cases where the correlation is strong under both hypotheses, we have observed that g_2 and g_3 perform better than the optimal iid nonlinearity. S_2 appears to be a slightly better performance measure than S_3; however, S_3 is much more mathematically tractable, since it usually leads to linear problems while S_2 leads to nonlinear problems. Of all the performance measures, S_3 appears to us to be the most intuitively pleasing.

In the numerical results of Section 5, we chose the parameters of the densities to match the first and second moments. We did this in order to make the discrimination problem difficult. When the densities are not closely matched as such, we have found that the iid nonlinearity often performs better. Thus, although we are not working with a weak signal model, we have found that our results are best when we are "close" to a weak signal situation. Evidently, performance measures which resemble signal-to-noise ratios are most useful for problems in which the marginal densities are not too dissimilar.

A final comment concerns the correlation we have observed between the theory, as presented in the first four sections of this paper, and the numerical results we obtained in applications. The performance measures are derived as asymptotic performance measures. Our numerical results, on the other hand, are obtained for finite sample sizes, and therefore do not illustrate the asymptotic performance as predicted in the theory. The derivations of the performance measures depended upon the assumption that the test statistic has a distribution which is approximately Gaussian when the sample size is large. Because the nonlinearities which optimize the performance measures S_0 and S_1 exhibit extremely large magnitudes near one of the endpoints of the region of integration,

the distribution of T_n in such cases is strongly skewed and the assumption of a Gaussian distribution is not valid. We have found that this need not be a problem, though, since the outlying values tend to fall away from the threshold. We have seen, in fact, that the nonlinearity which maximizes S_1 performs best relative to the other nonlinearities in certain cases. The other performance measures yield nonlinearities which are more "balanced," and consequently the distributions of the test statistics are more closely correlated to that predicted by the theory. According to our analytical results, it is possible to improve on the performance of the iid nonlinearity asymptotically. Our numerical results have demonstrated that a significant improvement in performance can be obtained even for relatively small sample sizes.

Appendix A: Mixing processes

To make mathematically precise the concept of asymptotic independence, we may define conditions referred to as (strong) mixing conditions. In this appendix, we state definitions and results which are useful for our work. Proofs of the results are omitted. The recent survey papers, [19] and [20], are excellent references.

Let $\{X_i, i \in \mathbf{Z}\}$ be a stationary stochastic sequence. Our results hold with an obvious modification for one-sided sided sequences. Let $\mathcal{F}_a^b$ be the σ-field generated by $\{X_i, a \leq i \leq b\}$ where a may take any finite value or the value $-\infty$, and similarly, b can take any value larger than or equal to a, including $+\infty$. Define now the following sequences:

$$\alpha_n = \sup_{r \in \mathbf{Z}} \{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^r, B \in \mathcal{F}_{r+n}^{\infty}\}$$

$$\rho_n = \sup_{r \in \mathbf{Z}} \{|\mathrm{Corr}(Y, Z)| : Y \in \mathcal{L}_2(\mathcal{F}_{-\infty}^r), Z \in \mathcal{L}_2(\mathcal{F}_{r+n}^{\infty})\}$$

$$\phi_n = \sup_{r \in \mathbf{Z}} \{|P(B|A) - P(B)| : A \in \mathcal{F}_{-\infty}^r, B \in \mathcal{F}_{r+n}^{\infty}\}$$

We say that the stochastic process $\{X_n\}$ is

strong mixing (or α-mixing) if $\alpha_n \to 0$ as $n \to 0$,

ρ-mixing if $\rho_n \to 0$ as $n \to \infty$,

ϕ-mixing if $\phi_n \to 0$ as $n \to \infty$,

m-dependent if $\phi_n = 0$ for $n > m$.

The list above is in the order of decreasing generality. Thus we have

$$m\text{-dependence} \Longrightarrow \phi\text{-mixing} \Longrightarrow \rho\text{-mixing} \Longrightarrow \text{strong mixing.}$$

The definitions above apply to nonstationary sequences. If the processes are stationary, then the supremum in each case need not be taken over r.

Any of the above mixing conditions are sufficient to imply that a stationary process is ergodic.

Because mixing conditions are defined on the σ-fields generated by the process $\{X_i\}$, it follows that for any measurable function g, the process $\{Y_i\}$ defined by $Y_i = g(X_i)$ will satisfy the same mixing conditions as $\{X_i\}$. Therefore, such conditions are ideal for obtaining results for memoryless nonlinearities.

A vast variety of central limit theorems have been proved for mixing processes. Generally, the strength of the auxilary conditions required depends on the relative strength of the mixing condition involved. For example, weaker auxilary conditions are needed for ϕ-mixing processes than for strong mixing ones. Two types of CLT's are possible. One type concludes that the sum S_n, when centered and normalized, converges in distribution to a standard normal random variable. A stronger type, which implies the former type, concludes that $S_{[nt]}$, when centered and normalized, converges in distribution to standard Brownian motion for $0 \leq t \leq 1$. The latter type is often called a functional central limit theorem. The following is such a theorem.

Theorem A *Let $\{X_n\}$ be a strictly stationary, zero-mean, strong mixing sequence, and $S_n = \sum_{k=1}^{n} X_k$. Let $\sigma_n^2 = \operatorname{Var} S_n$ and let $[\cdot]$ denote the greatest integer function. If $\sigma_n \to \infty$ and one of the following two conditions hold:*

(1) for some $\delta > 0$, $\mathrm{E}|X_1|^{2+\delta} < \infty$ *and* $\sum_n n\alpha(n)^{\delta/(2+\delta)} < \infty$

(2) for some $C > 0$, $|X_1| < C$ *and* $\sum_n n\alpha(n) < \infty$

then $S_{[nt]}/\sigma_n$ *converges in distribution to standard Brownian motion on* $[0, 1]$.

Appendix B: On the asymptotic normality of the test statistic for lognormal and Rice densities

We assume that the processes $\{Y_i\}$ and $\{Z_i\}$ are independent, stationary Gaussian random processes with mean $\tilde{\mu}$, variance $\tilde{\sigma}^2$, and covariances given by $\mathrm{Cov}[Y_1, Y_{j+1}] = \mathrm{Cov}[Z_1, Z_{j+1}] = \tilde{\rho}^j\tilde{\sigma}^2$. We consider two cases for the observation process $\{X_i\}$

(1) $X_i = \exp(Y_i)$.

(2) $X_i = \sqrt{Y_i^2 + Z_i^2}$.

In case (1), $\{X_i\}$ has a lognormal distribution. In case (2), $\{X_i\}$ has a Rayleigh distribution if $\tilde{\mu} = 0$, and a Rice distribution otherwise. The particular value of $\tilde{\sigma}^2$ is not relevant to our proof, so without loss of generality we assume $\tilde{\sigma}^2 = 1$. We require that $0 < \tilde{\rho} < 1$. The tildes on the parameters are necessary to distinguish them from μ and σ^2 defined by (1.4) and (1.5) and from the ρ parameters which define the ρ-mixing condition. It can be shown that the Gaussian processes $\{Y_i\}$ and $\{Z_i\}$ are Markov processes.

We wish to show that the test statistic T_n defined in (1.2) is such that $n^{-1/2}[T_n - n\mu]/\sigma$ converges in distribution as $n \to \infty$ to a standard normal random variable W (which has zero mean and unit variance). The particular central limit theorem which we will use is Theorem A. In this appendix, we show that the conditions of the theorem are met for the observation processes of cases (1) and (2) above. Statements are made without proof. Proofs for all statements may be found in [19] and [20].

The spectral density of the processes $\{Y_n\}$ and $\{Z_n\}$ is

$$S(\omega) = \frac{-e^{i\omega}(1-\tilde{\rho}^2)}{\tilde{\rho}(e^{i\omega})^2 - (1+\tilde{\rho}^2)e^{i\omega} + 1} = \frac{\tilde{\rho}\cos\omega - \tilde{\rho}^2}{1 - 2\tilde{\rho}\cos\omega + \tilde{\rho}^2}$$

The hypothesis that $0 < \rho < 1$ is sufficient to guarantee that the spectral density is positive for all ω values, and this is sufficient to imply that $\{Y_n\}$ and $\{Z_n\}$ are ρ mixing. Furthermore, the parameters ρ_n converge to 0 exponentially fast, since the processes are Markov. The bound $\alpha_n \leq \frac{1}{4}\rho_n$ implies the conditions (1) and (2) of Theorem A for the lognormal case. For the Rice (or Rayleigh) case, we have also the bound $\alpha_X < \alpha_Y + \alpha_Z$, which relates the parameters for $\{X_n\}$ and the parameters for $\{Y_n\}$ and $\{Z_n\}$. Thus for both the lognormal and the Rice processes, the conditions on the α parameters in Theorem A are satisfied. The other hypothesis for Theorem A, that σ_n converge to ∞, is a condition which we shall not be able to prove even in this very specific case. The case that σ_n does not approach ∞ is a degenerate case, and implies that $\sigma^2 = 0$ and hence that $n^{-1/2}[T_n - n\mu]$ converges in distribution to a constant. Otherwise, $n^{-1/2}[T_n - n\mu]/\sigma$ converges to a standard Gaussian.

Appendix C: Proof of Theorem 8

The proof of Theorem 8 for the performance measures S_0 and S_1 follows from our previous discussion concerning the performance measure $\hat{S}$ (see equation (4.6) and the discussion following) and the result published by Poor [16]. Therefore it remains to show that the theorem holds for S_3. We argued in Section IV that the problem reduces to one which involves only the marginal densities. Define the performance measure

$$\hat{S}_3(g, f_0, f_1) = \frac{[\mu(g; f_1) - \mu(g; f_0)]^2}{\sigma^2(g,; f_0) + A\sigma^2(g; f_1)}, \qquad (C.1)$$

where $A = [(1+2R_1)/(1+2R_0)]$. We must show that the least favorable marginal distributions f_0^* and f_1^* for the performance measure $\hat{S}_3$ are given by the Huber-Strassen densities (4.7).

A simple lemma which will be used is the following, whose proof is given in [13].

Lemma C. *If $v_1 > 0$, $v_2 > 0$, and $0 \leq \alpha \leq 1$ then*

$$\frac{\left[\alpha u_1 + (1-\alpha)u_2\right]^2}{\alpha v_1 + (1-\alpha)v_2} \leq \alpha \frac{u_1^2}{v_1} + (1-\alpha)\frac{u_2^2}{v_2}.$$

The integral equation which yields the optimal nonlinearity for $\hat{S}_3$ is similar to (3.19) with $m = 0$ except for the coefficient A:

$$g(x) = \frac{f_1(x) - f_0(x)}{f_0(x) + Af_1(x)} + \int \left[\frac{f_0(x)f_0(y) + Af_1(x)f_1(y)}{f_0(x) + Af_1(x)}\right] g(y)dy. \tag{C.2}$$

We have immediately the form of the solution

$$g(x) = \frac{B_0 f_0(x) + B_1 f_1(x)}{f_0(x) + Af_1(x)} \tag{C.3}$$

where $B_0 = \mu_0 - 1$ and $B_1 = A\mu_1 + 1$. If we consider the linear system of equations

$$\mu_0 = B_0 + 1 = \int g(x)f_0(x)dx$$

$$\mu_1 = \frac{1}{A}(B_1 - 1) = \int g(x)f_1(x)dx$$

with B_0, B_1 as the unknowns, then we find that the system is singular, and consequently we may assign to B_0 the arbtrary value 0. This implies that

$$B_1 = \left[\int \frac{f_0(x)f_1(x)}{f_0(x) + Af_1(x)}dx\right]^{-1}. \tag{C.4}$$

If g is the optimal nonlinearity which is matched to f_0, f_1, then we know that

$$\begin{aligned} \hat{S}_3(g, f_0, f_1) &= \int g(x)[f_1(x) - f_0(x)]dx \\ &= B_1(f_0, f_1) \int \frac{f_1(x)}{f_0(x) + Af_1(x)}[f_1(x) - f_0(x)]dx. \end{aligned} \tag{C.5}$$

where we have written B_1 as a function of f_0 and f_1 to remind us of the relation (C.4). Lemma 7 applies to the integral in (C.5) with $\Psi(x) = x(x-1)/(x+1)$, which is convex, so that the integral in (C.5) is minimized by the Huber-Strassen densities. Lemma 7 also applies to the integral in (C.4). In this case $\Psi(x) = x/(Ax+1)$, which is concave, so that by applying the lemma to the negative of the integral (since $-\Psi$ is convex) we find that this integral is maximized by the Huber-Strassen densities. $B_1(f_0, f_1)$ therefore is minimized. Thus our candidates for the least favorable marginal densities are the Huber-Strassen densities. The right inequality in (4.1) will now be proved.

The following inequalities, which depend on the fact that $\sigma^2(g; f)$ is concave in f and on Lemma C, demonstrate that $\hat{S}_3(g, f_0, f_1)$ is convex in f_0 and f_1 for fixed g. With $\beta = (1-\alpha)$ we have

$$\begin{aligned}
\hat{S}_3(g, \beta\bar{f}_0 + \alpha f_0, \beta\bar{f}_1 + \alpha f_0) &= \frac{\left[\mu(g; \beta\bar{f}_1 + \alpha f_1) - \mu(g; \beta\bar{f}_0 + \alpha f_0)\right]^2}{\sigma^2(g; \beta\bar{f}_0 + \alpha f_0) + A\sigma^2(g; \beta\bar{f}_1 + \alpha f_1)} \\
&\leq \frac{\left[\beta\{\mu(g; \bar{f}_1) - \mu(g; \bar{f}_0)\} + \alpha\{\mu(g; f_1) - \mu(g; f_0)\}\right]^2}{\beta\left[\sigma^2(g; \bar{f}_0) + A\sigma^2(g; \bar{f}_1)\right] + \alpha\left[\sigma^2(g; f_0) + A\sigma^2(g; f_1)\right]} \\
&\leq \beta\hat{S}_3(g, \bar{f}_0, \bar{f}_1) + \alpha\hat{S}_3(g, f_0, f_1)
\end{aligned}$$

Define the function

$$J(\alpha; f_0, f_1) = \hat{S}_3\left[g^*, (1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1\right] \qquad 0 \leq \alpha \leq 1$$

where f_0^* and f_1^* are the Huber-Strassen least favorable densities and g^* is the optimal nonlinearity matched to f_0^*, f_1^*. Certainly J is convex in α if $\hat{S}_3$ is convex in f_0 and f_1. Now the right inequality in (4.1) holds if and only if

$$J(\alpha; f_0, f_1) \geq J(0; f_0, f_1) \tag{C.6}$$

for all α in the interval $[0, 1]$, and since J is convex in α, (C.6) holds if and only if we have the condition

$$\frac{d}{d\alpha} J(\alpha; f_0, f_1)\Big|_{\alpha=0} \geq 0. \tag{C.7}$$

If we take the derivative of $J(\alpha; f_0, f_1)$ and set $\alpha = 0$, then we have

$$\begin{aligned}\frac{d}{d\alpha}J(\alpha; f_0, f_1)\Big|_{\alpha=0} &= 2\int g^*(f_1 - f_0 - f_1^* + f_0^*) - \int (g^*)^2(f_0 + Af_1) + \int (g^*)^2(f_0^* + Af_1^*) \\ &\quad + 2\Big[\int g^* f_0^*\Big]\Big[\int g^*(f_0 - f_0^*)\Big] + 2A\Big[\int g^* f_1^*\Big]\Big[\int g^*(f_1 - f_1^*)\Big] \qquad (C.8) \\ &= 2B_1\Big[\int g^* f_1 - \int g^* f_1^*\Big] + \int (g^*)^2(f_0^* + Af_1^*) - \int (g^*)^2(f_0 + Af_1).\end{aligned}$$

We can now show that (C.7) holds by considering the function

$$T[f_0, f_1] = \int \frac{f_1(x)^2}{f_0(x) + Af_1(x)}dx \qquad (C.9)$$

which by Lemma 7 is minimized by the Huber-Strassen densities. Define

$$K(\alpha; f_0, f_1) = T\big[(1-\alpha)f_0^* + \alpha f_0, (1-\alpha)f_1^* + \alpha f_1\big].$$

It follows from Lemma C that the integrand in (C.9) is convex in f_0, f_1, and thus T is convex. By the same reasoning as before, then, we conclude that f_0^*, f_1^* minimize T if and only if

$$\frac{d}{d\alpha}K(\alpha; f_0, f_1)\Big|_{\alpha=0} \geq 0. \qquad (C.10)$$

We now proceed to show that the inequality (C.10) implies the inequality (C.7).

Define $p_\alpha^{(i)} = (1-\alpha)f_i^* + \alpha f_i$ for $i = 0, 1$ and $0 \leq \alpha \leq 1$, so that we have

$$K(\alpha; f_0, f_1) = \int \frac{(p_\alpha^{(1)})^2}{p_\alpha^{(0)} + Ap_\alpha^{(1)}}. \qquad (C.11)$$

To show that we can interchange differentiation end integration to obtain an expression for $\frac{d}{d\alpha}K(\alpha; f_0, f_1)\big|_{\alpha=0}$, we write

$$\begin{aligned}\frac{d}{d\alpha}\left[\frac{(p_\alpha^1)^2}{p_\alpha^0 + Ap_\alpha^1}\right]_{\alpha=0} &\leq \frac{1}{\alpha}\left[\frac{(p_\alpha^{(1)})^2}{p_\alpha^{(0)} + Ap_\alpha^{(1)}} - \frac{(p_0^{(1)})^2}{p_0^{(0)} + Ap_0^{(1)}}\right] \\ &\leq \frac{(p_1^{(1)})^2}{p_1^{(0)} + Ap_1^{(1)}} - \frac{(p_0^{(1)})^2}{p_0^{(0)} + Ap_0^{(1)}}.\end{aligned} \qquad (C.12)$$

The inequalities in (C.12) are justified by the convexity of the integrand in (C.11) as a function of α. The right quantity in (C.12) is integrable, and the middle quantity converges pointwise monotonically to the left quantity as $\alpha \to 0$ because of the convexity of the integrand in (C.11). The monotone convergence theorem then permits the interchange of the differentiation and the integration, and we have

$$\frac{d}{d\alpha} K(\alpha; f_0, f_1)\Big|_{\alpha=0} =$$

$$\int \left\{ 2\frac{f_1^*}{f_0^* + Af_1^*}(f_1 - f_1^*) + \left(\frac{f_1^*}{f_0^* + Af_1^*}\right)^2 [(f_0^* + Af_1^*) - (f_0 + Af_1)] \right\}. \qquad (C.13)$$

Now if we compare equations (C.8) and (C.13), then we see that (compare (C.3))

$$\frac{d}{d\alpha} J(\alpha; f_0, f_1)\Big|_{\alpha=0} = \frac{d}{d\alpha} B_1^2 K(\alpha; f_0, f_1)\Big|_{\alpha=0}$$

and thus conditions (C.7) and (C.10) are equivalent.

References

[1] H. V. Poor and J. B. Thomas, "Memoryless Discrete-Time Detection of a Constant Signal in m-Dependent Noise," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 54–61, Jan. 1979.

[2] D. R. Halverson and G. L. Wise, "Discrete-Time Detection in ϕ-Mixing Noise," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 189–198, Mar. 1980.

[3] J. S. Sadowsky and J. A. Bucklew, "A Nonlocal Approach for Asymptotic Memoryless Detection Theory," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 115–120, Jan. 1986.

[4] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968, pp. 166–195.

[5] J. F. Barrett and D. G. Lampard, "An Expansion For Some Second-Order Probability Distributions and Its Application to Noise Problems," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 10–15, March 1955.

[6] G. E. Noether, "On a Theorem of Pitman," *Ann. Math Stat.*, vol. 26, pp.64–68, 1955.

[7] P. J. Huber, "A Robust Version of the Probability Ratio Test," *Ann. Math. Statist.* vol. 36, pp. 1753–1758.

[9] J. S. Sadowsky, "A Maximum Variance Model for Robust Detection and Estimation with Dependent Data," *IEEE Trans. Inform. Theory,* vol. IT-32, pp. 220–226, March 1986.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag: New York, 1988.

[11] J. H. Miller and J. B. Thomas, "Detectors for Discrete-Time Signals in Non-Gaussian Noise," *IEEE Trans. Inform. Theory,* vol. IT-18, pp. 241–250, March 1972.

[12] G. V. Moustakides and J. B. Thomas, "Optimum Detection of a Weak Signal with a Minimal Knowledge of Dependency," *IEEE Trans. Inform. Theory,* vol. IT-32, pp. 97–102, Jan. 1986.

[13] P. J. Huber, "Robust Estimation of a Location Parameter," *Ann. Math. Statist.* vol. 35, pp. 73–101.

[14] P. J. Huber and V. Strassen, "Minimax Tests and the Neyman-Pearson Lemma for Capacities," *Ann. Stat.,* vol. 1, pp. 251–263, 1973.

[15] H. V. Poor, "On Robust Wiener Filtering," *IEEE Trans. Auto. Control* vol. AC-25, pp. 531–536, June 1980.

[16] H. V. Poor, "Robust Decision Design Using a Distance Criterion," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 575–587, September 1980.

[17] S. A. Kassam, "Robust Hypothesis Testing For Bounded Classes of Probability Densities," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 242–247, 1981.

[18] K. S. Vastola and H. V. Poor, "On the p-point Uncertainty Class," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 374–376, 1984.

[19] R. C. Bradley, "The Basic Properties of Strong Mixing Conditions," in *Dependence in Probability and Statistics*, Birkhauser, 1985.

[20] M. Peligrad, "Recent Advances in the Central Limit Theorem and Its Weak Invariance

Principle for Mixing Sequences of Random Variables," in *Dependence in Probability and Statistics*, Birkhauser, 1985.

[21] B. Picinbono and P. Duvaut, "Optimal Linear-Quadratic Systems for Detection and Estimation," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 304–311, March 1988.

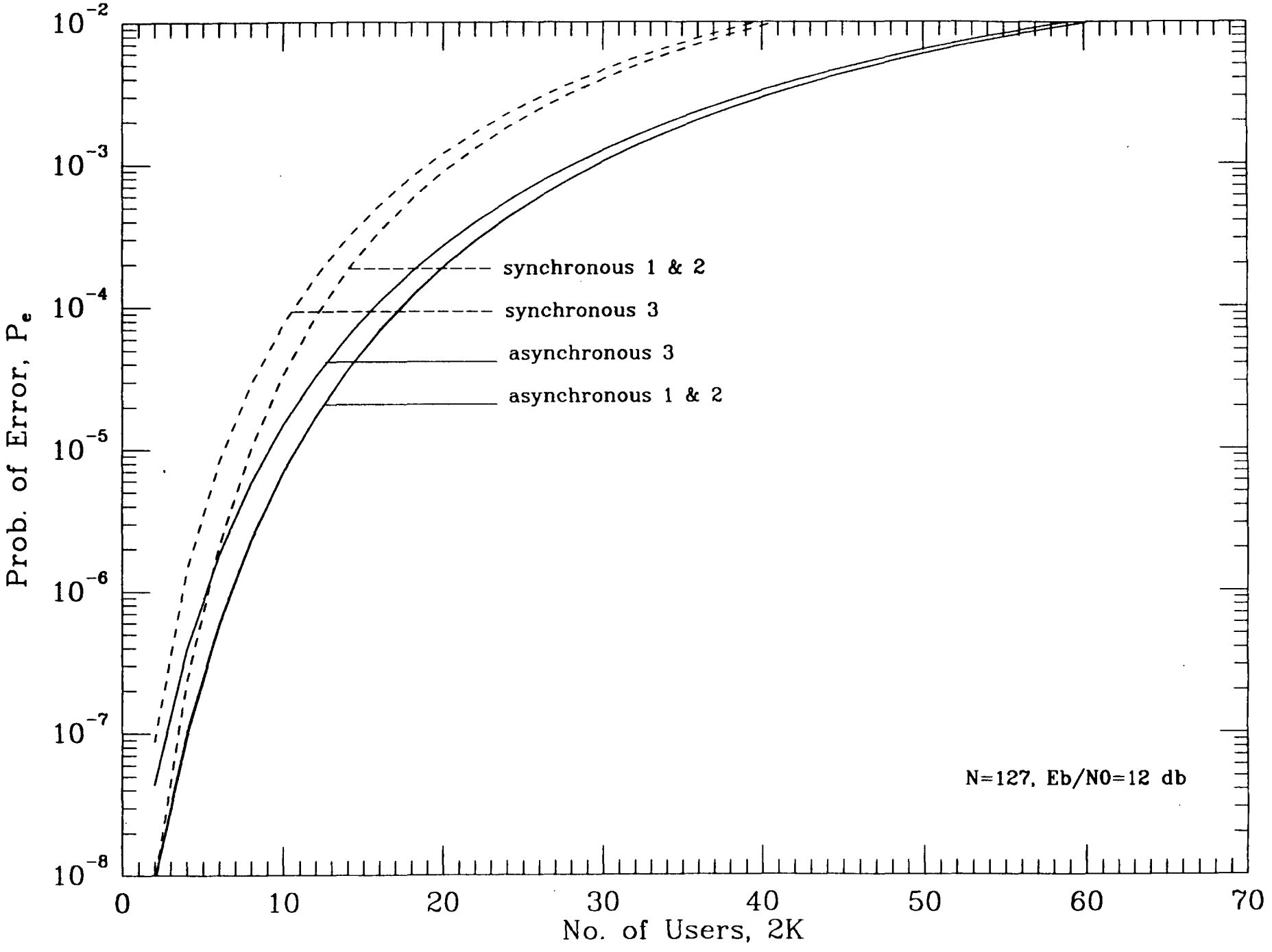

*Fig.*1. Broadcast capability of DS systems.

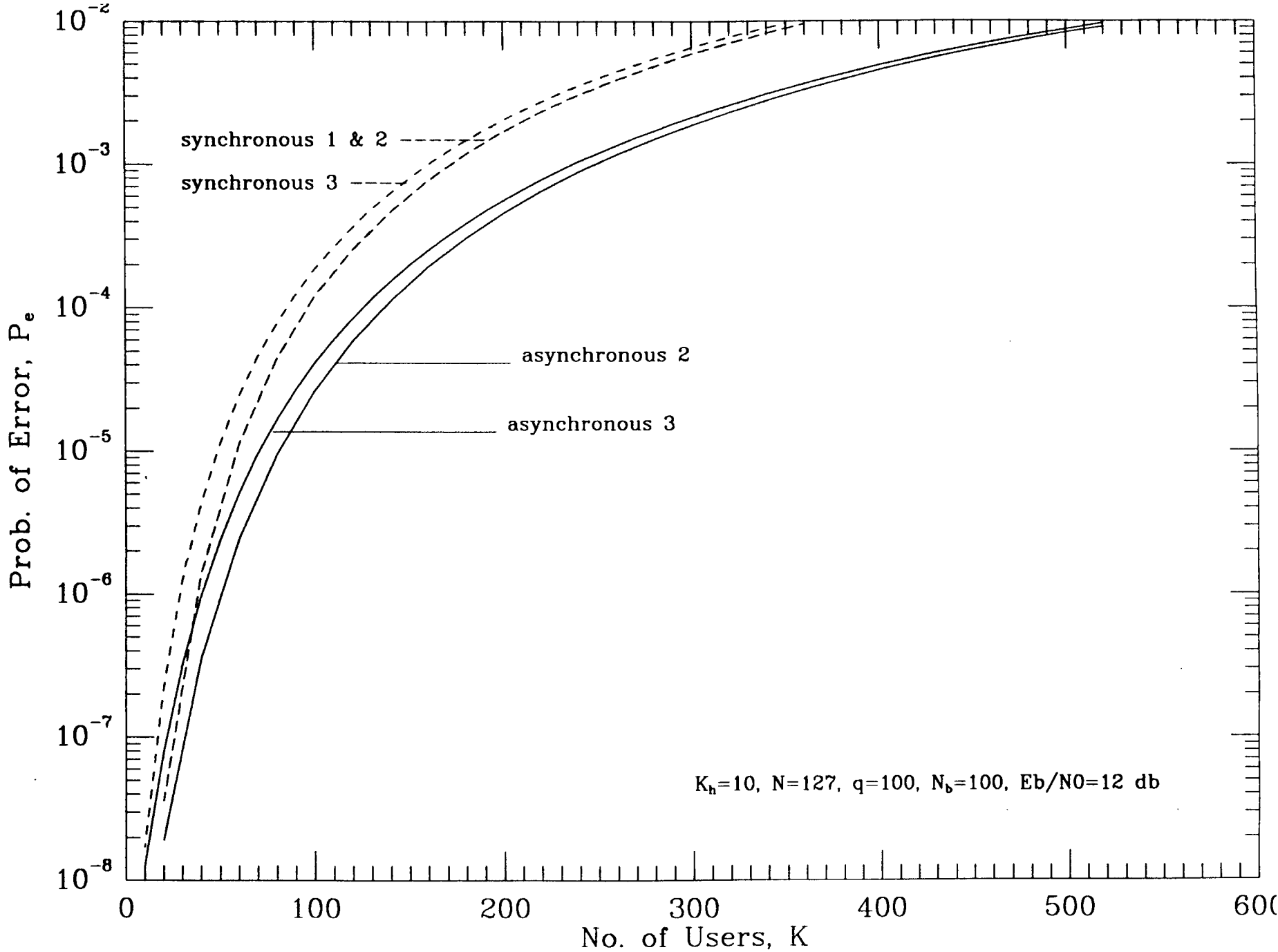

*Fig.*2. Broadcast capability of **hybrid** systems for $K_h = 10$.

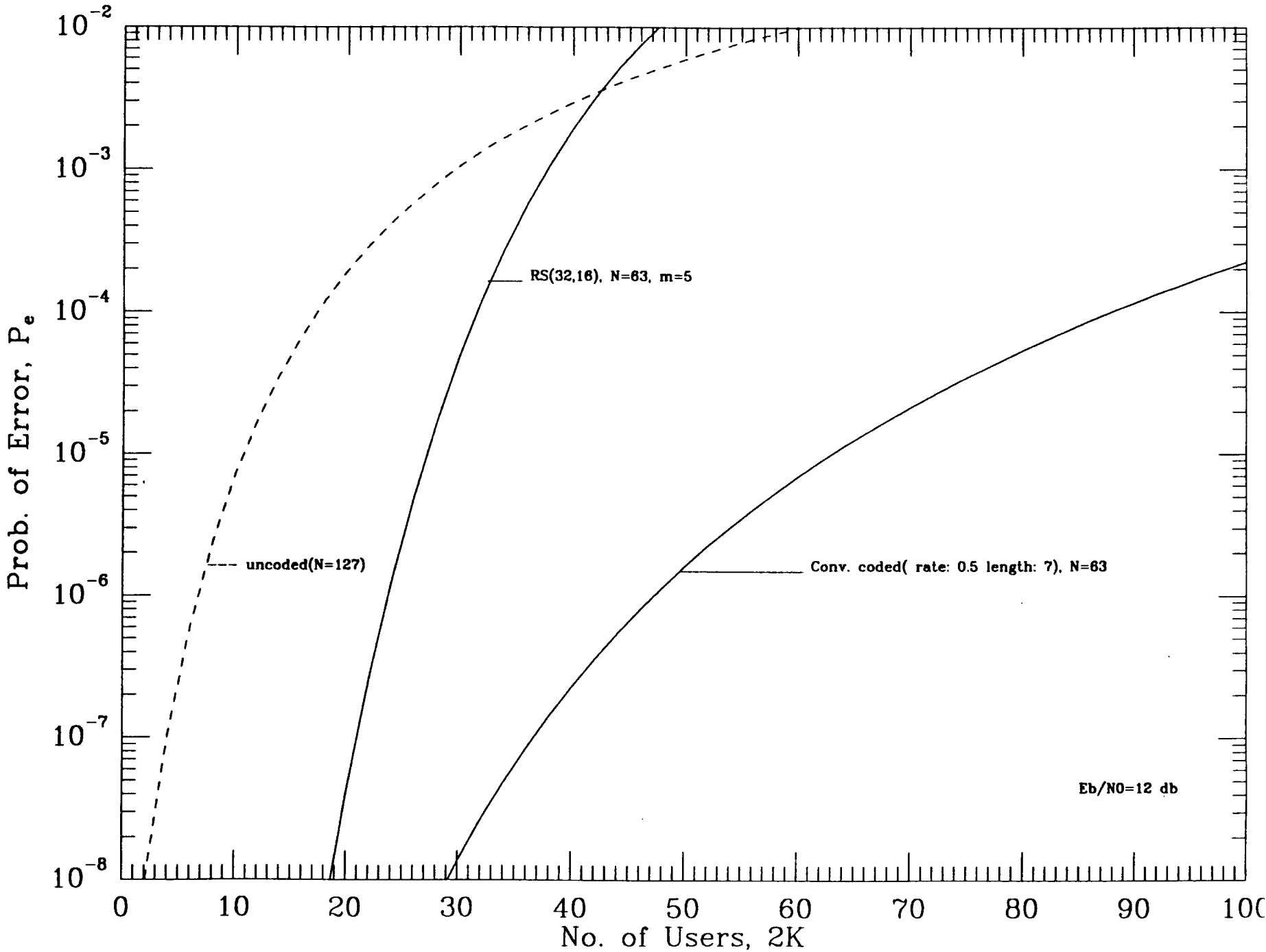

*Fig.*3. Broadcast capability of asynchronous DS System 2 for uncoded, convolutional and RS coded systems.

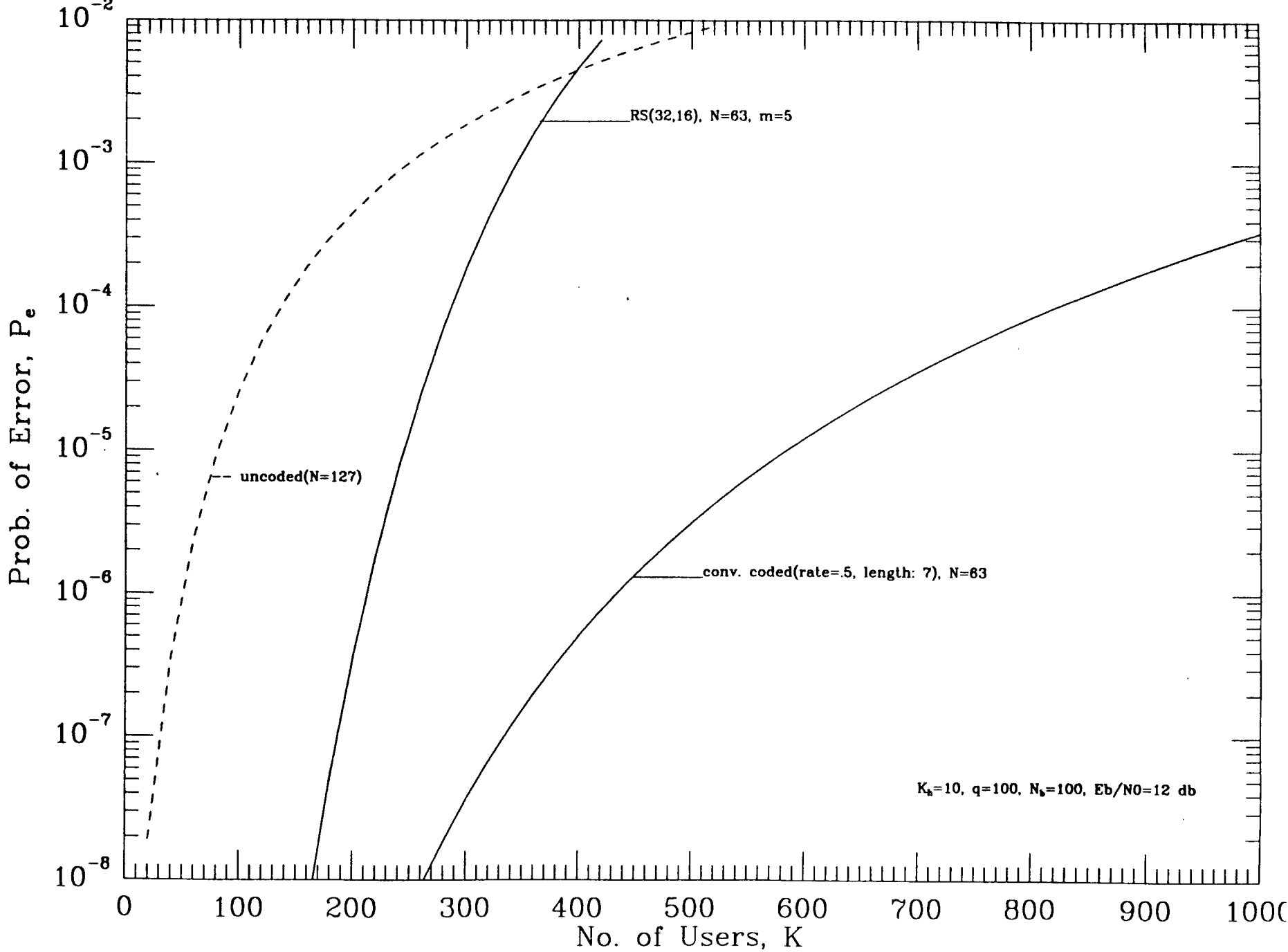

*Fig.*4. Broadcast capability of asynchronous hybrid System 2 for uncoded, convolutional and RS coded systems with $K_h = 10$.