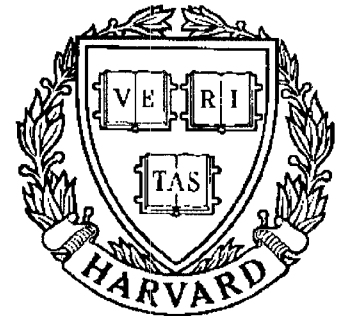


TECHNICAL RESEARCH REPORT



S Y S T E M S
R E S E A R C H
C E N T E R



*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
the University of Maryland,
Harvard University,
and Industry*

Error Exponents for Distributed Detection of Markov Sources

by H.M.H. Shalaby and A. Papamarcou

ERROR EXPONENTS FOR DISTRIBUTED DETECTION OF MARKOV SOURCES¹

Hossam M. H. Shalaby and Adrian Papamarcou

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

Abstract

We consider a decentralized detection problem in which two sensors collect data from a discrete-time finite-valued stationary ergodic Markov source and transmit M -ary messages to a Neyman-Pearson central detector. We assume that the codebook sizes M are fixed for both sensors and do not vary with data sample size. We investigate the asymptotic behavior of the type II error rate as the sample size increases to infinity and obtain (under mild assumptions on the source distributions) the associated error exponent. The derived exponent is independent of the test level ϵ and the codebook sizes M , is achieved by a universally optimal sequence of acceptance regions and is characterized by an infimum of informational divergence over a class of infinite-dimensional distributions.

1. Problem statement and background

In this paper we discuss the asymptotically optimal design of a distributed detection system for stationary ergodic Markov sources. Our setup comprises

- (i) a discrete-time information source $(X_i, Y_i)_{i=-\infty}^{\infty}$,
- (ii) two remote sensors S_X and S_Y ;
- (iii) a central detector or fusion center.

The source is assumed to be stationary ergodic Markov with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, where \mathcal{X} and \mathcal{Y} are finite. The sensors S_X and S_Y observe the sequences X_1^n and Y_1^n respectively, and encode their observations into single messages taking one of $M_X \geq 2$ and $M_Y \geq 2$ values, respectively. It should be emphasized that the codebook sizes (or equivalently, quantization levels) M_X and M_Y are fixed in n . These messages are communicated to the central detector, which proceeds to declare which of two hypotheses (H_0 or H_1) concerning the source statistics is true.

A classical (Neyman-Pearson) procedure for testing H_0 versus H_1 is assumed throughout, and our object is to study the asymptotic performance of the optimal test of level

1. Research supported by the Systems Research Center (a National Science Foundation Engineering Research Center) at the University of Maryland and by the Minta Martin Fund for Aeronautical Research administered by the College of Engineering, University of Maryland.

This paper was presented in part at the 1991 Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore MD, March 20-24, 1991.

$\epsilon \in (0, 1)$ based on n consecutive sensor observations. Specifically, if $\beta_n(M_X, M_Y, \epsilon)$ is the type II error probability of the above optimal test, we seek to determine the error exponent

$$\theta(M_X, M_Y, \epsilon) \stackrel{\text{def}}{=} - \limsup_n \frac{1}{n} \log \beta_n(M_X, M_Y, \epsilon) .$$

The codebook sizes M_X and M_Y are essential parameters in the above formulation, as they characterize the distributed nature of the detection process. Setting both M_X and M_Y equal to infinity yields the degenerate case of the conventional centralized detector, for which the error exponent is known [1]. Indeed, if $W(\cdot|\cdot)$ (resp. $V(\cdot|\cdot)$) is the Markov transition matrix of the source under H_0 (resp. H_1) and π_W is the stationary distribution (on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$) of the source under H_0 , then

$$\theta(\infty, \infty, \epsilon) = D(W||V|\pi_W) , \quad (1.1)$$

where $D(W||V|\pi_W)$ is the conditional informational divergence:

$$D(W||V|\pi_W) \stackrel{\text{def}}{=} \sum_{(z_1, z_2) \in \mathcal{Z}^2} \pi_W(z_1) W(z_2|z_1) \log \frac{W(z_2|z_1)}{V(z_2|z_1)} .$$

It should be noted that the expression for the error exponent in equation (1.1) does not involve ϵ and depends on the source distribution only through its restriction to two consecutive time coordinates.

Our search for $\theta(M_X, M_Y, \epsilon)$ in the nondegenerate case follows earlier work [3,4] on distributed detection for memoryless sources exhibiting spatial dependence (such sources constitute a proper subclass of the Markov processes considered in this paper). These studies showed that if $(X_i, Y_i)_{i=1}^\infty$ is an i.i.d. process whose distribution is given by the infinite product of a bivariate P_{XY} (under H_0) or a bivariate $Q_{XY} > 0$ (under H_1) on $\mathcal{X} \times \mathcal{Y}$, then the error exponent is given by

$$\theta_{\text{iid}}(M_X, M_Y, \epsilon) = \min_{\tilde{P}_{XY}: \tilde{P}_X = P_X, \tilde{P}_Y = P_Y} D(\tilde{P}_{XY}||Q_{XY}) \quad (1.2)$$

provided at least one of M_X and M_Y is finite. Here $D(\cdot||\cdot)$ is the usual (not conditional) informational divergence, defined by

$$D(P||Q) \stackrel{\text{def}}{=} \sum_{z \in \mathcal{Z}} P(z) \log \frac{P(z)}{Q(z)} .$$

The expression for the error exponent in (2) does not involve M_X , M_Y or ϵ , and depends on the source distribution only through its restriction to a single time coordinate.

2. Main contribution

Extrapolating from equations (1.1) and (1.2) of the previous section, one might conjecture that for the nondegenerate Markov problem

- (i) the error exponent $\theta(M_X, M_Y, \epsilon)$ is independent of M_X , M_Y and ϵ ;
- (ii) a characterization of $\theta(M_X, M_Y, \epsilon)$ can be given in terms of the infimum of a suitable divergence functional over a computable class of distributions on $(\mathcal{X} \times \mathcal{Y})^2$.

Our main result indicates that only the first of the above two statements is true. Specifically, we prove the following.

Theorem 2.1. *If the alternative transition matrix $V(\cdot|\cdot)$ satisfies the positivity constraint $V(\cdot|\cdot) > 0$ and at least one of M_X, M_Y is finite, then the error exponent $\theta(M_X, M_Y, \epsilon)$ is given by the infimum of the conditional divergence*

$$E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}}$$

over all distributions \tilde{P} on $(\mathcal{X} \times \mathcal{Y})^{\mathbb{Z}}$ whose restrictions on $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{Y}^{\mathbb{Z}}$ agree with those of the null Markov distribution.

The above complete characterization of $\theta(M_X, M_Y, \epsilon)$ clearly involves a class of distributions on an infinite-dimensional space and does not appear to admit a reduction to a finite dimension. Thus statement (ii) seems to be false.

In the remainder of the paper, we give the principal arguments in the proof of Theorem 2.1, and indicate possible extensions to more general systems. Technical preliminaries appear in Section 3. The direct (or positive) part of Theorem 2.1 is established in Section 4, and the converse in Section 5. Concluding remarks are made in Section 6.

3. Preliminaries.

(a) General notation. For simplicity we let $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{Y}$ and $Z_i \stackrel{\text{def}}{=} (X_i, Y_i)$. As usual, Z_i^j and z_i^j denote (Z_i, \dots, Z_j) and (z_i, \dots, z_j) , respectively.

The symbols W_k and V_k denote transition matrices for $(k-1)$ -order Markov processes. Thus $W_k(\cdot|\cdot)$ is a stochastic matrix on $\mathcal{Z} \times \mathcal{Z}^{k-1}$ and $W_k(z_k|z_1^{k-1})$ is the probability that the corresponding process produces the symbol z_k given a preceding string $z_1^{k-1} = (z_1, \dots, z_{k-1})$.

Given a stochastic matrix W_k on $\mathcal{Z} \times \mathcal{Z}^{k-1}$ which induces a unique stationary distribution π_W on \mathcal{Z}^{k-1} , we let $P = \pi_W \diamond W_k$ denote the associated $(k-1)$ -order Markov extension on the Borel field of $\mathcal{Z}^{\mathbb{Z}}$. The measure P is stationary; if W_k is aperiodic, P is also ergodic. The entropy rate of $P = \pi_W \diamond W_k$ is defined by

$$H(W_k|\pi_W) = - \sum_{z_1^k \in \mathcal{Z}^k} \pi_W(z_1^{k-1}) W_k(z_k|z_1^{k-1}) \log W_k(z_k|z_1^{k-1})$$

and for $l \leq k$, the divergence rate $D(W_k || V_l | \pi_W)$ is defined by

$$D(W_k || V_l | \pi_W) \stackrel{\text{def}}{=} \sum_{z_1^k \in \mathcal{Z}^k} \pi_W(z_1^{k-1}) W_k(z_k | z_1^{k-1}) \log \frac{W_k(z_k | z_1^{k-1})}{V_l(z_k | z_{k-l+1}^{k-1})}.$$

(b) Acceptance regions. As stated in Section 1, sensors S_X and S_Y encode their observations X_1^n and Y_1^n into one of M_X and M_Y messages, respectively. Thus for any given n , S_X partitions the space \mathcal{X}^n into M_X cells, denoted by $C_n^{(1)}, \dots, C_n^{(M_X)}$; and S_Y partitions \mathcal{Y}^n into M_Y cells, denoted by $G_n^{(1)}, \dots, G_n^{(M_Y)}$. Each sensor then communicates to the central detector the cell index corresponding to its observation. This constrains the central detector to employ an acceptance region (for the null hypothesis) of the form

$$\mathcal{A}_n = \bigcup_{i=1}^{M_X} C_n^{(i)} \times F_n^{(i)}, \quad (3.1)$$

where each $F_n^{(i)}$ is a possibly empty union of cells $G_n^{(j)}$.

(c) Typical sequences. Typical sequences are commonly employed in proving direct results in Shannon theory, as they provide a convenient means of identifying sets of high probability under i.i.d. or Markov measures that are also in a certain sense “lean,” or non-redundant. We give here a summary of pertinent facts on the Markov concept of typicality.

The k -order type of a sequence $z_1^n \in \mathcal{Z}^n$ is the empirical distribution $\lambda_{z_1^n}^k$, or simply λ_z^k , on \mathcal{Z}^k resulting from computing the relative frequency of each k -string along the periodic extension of z_1^n . This method of evaluation of relative frequency ensures that the marginals of λ_z^k are translation-invariant. That is to say, if both I and $I+1$ are subsets of $\{1, \dots, k\}$, then the marginals of λ_z^k corresponding to the index sets I and $I+1$ are identical.

Given any k -order type \hat{P}^k , we can define a distribution $\hat{\pi}$ on \mathcal{Z}^{k-1} and a $(k-1)$ -order Markov transition matrix \hat{W}_k on $\mathcal{Z} \times \mathcal{Z}^{k-1}$ by

$$\hat{\pi}_W(z_1^{k-1}) \stackrel{\text{def}}{=} \sum_{z_k \in \mathcal{Z}} \hat{P}_k(z_1^k) \quad (3.2)$$

and

$$\hat{W}_k(z_k | z_1^{k-1}) \stackrel{\text{def}}{=} \begin{cases} \hat{P}_k(z_1^k) / \hat{\pi}_W(z_1^{k-1}), & \text{if } \hat{\pi}_W(z_1^{k-1}) > 0; \\ 1/|\mathcal{Z}|, & \text{otherwise.} \end{cases} \quad (3.3)$$

It is easy to see that $\hat{\pi}$ is a stationary distribution for \hat{W}_k . Also, the definition of k -order type in terms of relative frequency along a periodic sequence implies the existence of a single recurrent class in \hat{W}_k . Hence $\hat{\pi}$ is the unique stationary distribution for \hat{W}_k , and we may write $\hat{\pi} = \pi_{\hat{W}}$. In addition, we observe that \hat{P}^k is the restriction of the stationary measure $\hat{P} = \hat{\pi}_W \diamond \hat{W}_k$ to k successive time coordinates.

We denote the set of all k -order types obtained from sequences of length n by $\mathcal{P}_{(n)}^k(\mathcal{Z})$, and we distinguish between two notions of typicality. The first is precise: if $\hat{P}^k \in \mathcal{P}_{(n)}^k(\mathcal{Z})$, we say that a sequence z_1^n is \hat{P}^k -typical if $\lambda_z^k = \hat{P}^k$ and we denote the set of all such sequences by \hat{T}_Z^k . The second notion involves an approximation: if \tilde{P} is an arbitrary stationary distribution on $\mathcal{Z}^{\mathbb{Z}}$, we say that a sequence z_1^n is (\tilde{P}^k, η) -typical if

$$\sup_{a_1^k} |\lambda_z^k(a_1^k) - \tilde{P}(a_1^k)| \leq \eta$$

and we denote the set of all such sequences by $\tilde{T}_{Z,\eta}^k$. Note that the sequence length n is omitted from both \hat{T}_Z^k and $\tilde{T}_{Z,\eta}^k$ for brevity.

We give some standard facts on the cardinalities (denoted by $|\cdot|$) and probabilities of some of the sets introduced above in the following lemma. For a proof of assertions (i) and (ii), see [5]. Assertion (iii) is easily established using the pointwise ergodic theorem.

Lemma 3.2. (i) $|\mathcal{P}_{(n)}^k(\mathcal{Z})| \leq r(n)$, where $r(\cdot)$ is a polynomial of degree $|\mathcal{Z}|^{k-1}$.

(ii) Let $\hat{P}^k \in \mathcal{P}_{(n)}^k(\mathcal{Z})$ and $\hat{\pi}$, \hat{W}_k be as in (3.2) and (3.3). Then there exists a polynomial $s(\cdot)$ of degree $2|\mathcal{Z}|^k$ and an absolute constant c such that

$$[s(n)]^{-1} \exp[nH(\hat{W}_k|\hat{\pi})] \leq |\hat{T}_Z^k| \leq c \exp[nH(\hat{W}_k|\hat{\pi})] .$$

(iii) If \tilde{P} be a stationary ergodic measure on $\mathcal{Z}^{\mathbb{Z}}$, then for any $k \geq 1$ and $\eta > 0$, there exists a sequence $\{\xi_n\}_{n=1}^{\infty}$ with $\xi_n \rightarrow 0$ such that the set of (\tilde{P}^k, η) -typical sequences of length n satisfies

$$\tilde{P}(\tilde{T}_{Z,\eta}^k) \geq 1 - \xi_n .$$

(d) **Blowing-up lemma.** The blowing-up lemma of Ahlswede-Gács-Körner [2] is a powerful tool for proving converse theorems involving i.i.d. sources and was used in establishing the error exponent for distributed detection of such sources under zero-rate data compression [4]. In this paper we employ a version of the blowing-up lemma suitable for stationary ergodic Markov sources. It is derived from an ergodic theoretic result given in Rudolph [8] upon application of a technique which is also used in Marton [9].

Lemma 3.2. Let P be a strongly mixing stationary Markov measure on \mathcal{Z}^{∞} . If for fixed $\lambda > 0$ the set $B_n \subset \mathcal{Z}^n$ satisfies

$$(\forall n) \quad P(B_n) \geq \lambda ,$$

then for all k_n and sufficiently large n , the Hamming k -neighborhood of B_n —denoted by $\Gamma^{k_n} B_n$ —satisfies

$$P(\Gamma^{k_n} B_n) \geq 1 - \frac{n\lambda}{k_n} .$$

Although the above lemma is significantly weaker than its counterpart for i.i.d. measures, it suffices for the converse result obtained in this paper.

4. Direct theorem

We recapitulate the problem statement as follows. We are given a stationary ergodic Markov source $Z_{-\infty}^{\infty} = (X, Y)_{-\infty}^{\infty}$ and two simple hypotheses H_0 and H_1 . Under H_0 , the source distribution is given by $P = \pi_W \diamond W$, where $W = W_2$ is an irreducible aperiodic transition matrix. Under H_1 , we have $Q = \pi_V \diamond V$, where $V = V_2$ satisfies the additional positivity constraint $V > 0$. The optimal test of level ϵ based on n consecutive sensor observations is one that minimizes $Q(\mathcal{A}_n)$ (the probability of type II error) over all acceptance regions \mathcal{A}_n that

- yield a value of $P(\mathcal{A}_n^c)$ (probability of type I error) less than or equal to ϵ ; and
- satisfy condition (3.1).

The resulting *minimum* probability of type II error is denoted by $\beta_n(M_X, M_Y, \epsilon)$, and the error exponent $\theta(M_X, M_Y, \epsilon)$ is defined as in Section 1.

The positive theorem of this section yields a lower bound on the error exponent expressed in terms of linear subspaces \mathcal{L}^k and \mathcal{L} of distributions on $\mathcal{Z}^{\mathbb{Z}}$. These are defined by

$$\begin{aligned} \mathcal{L}^k &\stackrel{\text{def}}{=} \{ \tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbb{Z}} : \\ &\quad \tilde{P}(x_{-k+1}^0) = P(x_{-k+1}^0), \tilde{P}(y_{-k+1}^0) = P(y_{-k+1}^0) \} , \end{aligned} \quad (4.1)$$

$$\begin{aligned} \mathcal{L} &\stackrel{\text{def}}{=} \{ \tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbb{Z}} : \\ &\quad (\forall n \geq 0) \quad \tilde{P}(x_{-n}^0) = P(x_{-n}^0), \tilde{P}(y_{-n}^0) = P(y_{-n}^0) \} . \end{aligned} \quad (4.2)$$

It is clear that $\mathcal{L}^{k+1} \subset \mathcal{L}^k$ and $\mathcal{L}^k \searrow \mathcal{L}$.

Theorem 4.1. *If we let for $k \geq 2$*

$$D^k \stackrel{\text{def}}{=} \min_{\tilde{P} \in \mathcal{L}^k} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0 | (XY)_{-k+1}^{-1}\}}{V\{(XY)_0 | (XY)_{-1}\}} ,$$

then for all $\epsilon \in (0, 1)$ we have

$$\theta(2, 2, \epsilon) \geq \sup_k D^k \geq \inf_{\tilde{P} \in \mathcal{L}} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0 | (XY)_{-\infty}^{-1}\}}{V\{(XY)_0 | (XY)_{-1}\}} .$$

Proof. The idea is to construct for some k and $\eta > 0$ a sequence of acceptance regions $\mathcal{A}_n \subset \mathcal{Z}^n$ that contain the set $T_{Z, \eta}^k$ of (P^k, η) -typical sequences. By statement

(iii) of Lemma 3.2, this will ensure that $P(\mathcal{A}_n) \geq 1 - \epsilon$ for sufficiently large n , thereby satisfying the type I error constraint.

The set $T_{Z,\eta}^k$ itself cannot be expressed in the form of (3.1) for $M_X = 2$ and $M_Y = 2$ (or any other pair (M_X, M_Y) that is fixed in n), and thus the choice $\mathcal{A}_n = T_{Z,\eta}^k$ is not permissible. We consider instead the marginals P_X and P_Y of P on $\mathcal{X}^{\mathbf{Z}}$ and $\mathcal{Y}^{\mathbf{Z}}$ (respectively), and let

$$\mathcal{A}_n = T_{X,\eta}^k \times T_{Y,\eta}^k .$$

Here $T_{X,\eta}^k$ and $T_{Y,\eta}^k$ are the sets of (P_X^k, η) -typical sequences in \mathcal{X}^n and (P_Y^k, η) -typical sequences in \mathcal{Y}^n , respectively.

It is easy to establish that \mathcal{A}_n thus chosen satisfies the type I error constraint (for large n) and the compression constraint (3.1). To evaluate the corresponding type II error, we note that

$$\mathcal{A}_n = \bigcup_{\hat{P}^k \in \Phi^k} \hat{T}_Z^k , \quad (4.3)$$

where

$$\Phi^k = \{\hat{P} \in \mathcal{P}_{(n)}^k(\mathcal{X} \times \mathcal{Y}) : \sup |\hat{P}_X - P_X| \leq \eta, \sup |\hat{P}_Y - P_Y| \leq \eta\} \quad (4.4)$$

A routine computation based on Lemma 3.1 (ii) gives for any $\hat{P} \in \mathcal{P}_{(n)}^k(\mathcal{X} \times \mathcal{Y})$

$$\frac{1}{\sigma(n)} \exp[-nD(\hat{W}_k || V_l | \hat{\pi}_W)] \leq Q(\hat{T}_Z^k) \leq \gamma \exp[-nD(\hat{W}_k || V_l | \hat{\pi}_W)] \quad (4.5)$$

where $\sigma(n)$ is a polynomial of degree at most $2|\mathcal{Z}|^k$ and γ is an absolute constant. Equations (4.3), (4.4), (4.5) and the type counting result in Lemma 3.1 (i) together yield the estimate

$$\frac{1}{n} \log Q(\mathcal{A}_n) \leq - \min_{\hat{P}^k \in \Phi^k} D(\hat{W}_k || V | \hat{\pi}_W) + \delta_n ,$$

where $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. Continuity of the divergence functional enables us to approximate the above minimum over Φ^k by an infimum over \mathcal{L}^k , where \mathcal{L}^k is defined in (4.1). By compactness of \mathcal{L}^k , we can then replace the infimum by a minimum. We thus have for sufficiently small η ,

$$\frac{1}{n} \log Q(\mathcal{A}_n) \leq - \min_{\tilde{P} \in \mathcal{L}} \sum_{z_1^k \in \mathcal{Z}^k} \tilde{P}(z_1^k) \log \frac{\tilde{P}(z_k | z_1^{k-1})}{V(z_k | z_{k-1})} + \delta_n + \nu(\eta) ,$$

where $\nu(\eta) \rightarrow 0$ as $\eta \rightarrow 0$. Expressing the sum inside the minimum as an expectation (with a permissible index shift) and invoking the definition of the error exponent, we obtain

$$\begin{aligned} \theta(2, 2, \epsilon) &= - \limsup_n \frac{1}{n} \log \beta_n(2, 2, \epsilon) \\ &\geq - \limsup_n \frac{1}{n} \log Q(\mathcal{A}_n) \\ &= \min_{\tilde{P} \in \mathcal{L}^k} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0 | (XY)_{-k+1}^{-1}\}}{V\{(XY)_0 | (XY)_{-1}\}} = D^k . \end{aligned}$$

It follows that $\theta(2, 2, \epsilon) \geq \sup_k D^k$.

In order to complete the proof we need to show that

$$\lim_k D^k \geq \inf_{\tilde{P} \in \mathcal{L}} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}}. \quad (4.6)$$

To do so, we consider a sequence of measures $\{\mu_k\}$ on $\mathcal{Z}^{\mathbf{Z}}$ such that μ_k achieves the minimum in the definition of D_k . As the product space $\mathcal{Z}^{\mathbf{Z}}$ is equivalent to a compact metric space, we may invoke Prohorov's theorem [6] to conclude that $\{\mu_k\}$ contains a subsequence $\{\mu_{k_i}, i \in \mathbf{N}\}$ which converges weakly to a measure $\tilde{\mu}$. Thus every cylinder set $K \subset \mathcal{Z}^{\mathbf{Z}}$ —being both open and closed in the usual product topology—will satisfy

$$\lim_i \mu_{k_i}(K) = \tilde{\mu}(K). \quad (4.7)$$

It also follows that $\tilde{\mu}$ is stationary and lies in \mathcal{L} . It is not, however, necessarily ergodic. To establish (4.6), it suffices to show that

$$\lim_i D^{k_i} \geq \tilde{D},$$

where

$$\tilde{D} \stackrel{\text{def}}{=} E_{\tilde{\mu}} \log \frac{\tilde{\mu}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}}.$$

We do so in four steps:

Step 1. We approximate \tilde{D} by \tilde{D}_n , where

$$\tilde{D}_n \stackrel{\text{def}}{=} E_{\tilde{\mu}} \log \frac{\tilde{\mu}(Z_0|Z_{-n+1}^{-1})}{V(Z_0|Z_{-1})}.$$

We have

$$\tilde{D}_{n+1} - \tilde{D}_n = H_{\tilde{\mu}}(Z_0|Z_{-n+1}^{-1}) - H_{\tilde{\mu}}(Z_0|Z_{-n}^{-1}) \geq 0$$

and

$$\tilde{D} - \tilde{D}_n = H_{\tilde{\mu}}(Z_0|Z_{-n+1}^{-1}) - H_{\tilde{\mu}}(Z_0|Z_{-\infty}^{-1}) \geq 0.$$

By applying Levy's martingale convergence theorem (see also [7]), we obtain

$$\lim_n \tilde{D}_n = \sup_n \tilde{D}_n = \tilde{D}. \quad (4.8)$$

Step 2. We approximate \tilde{D}_n by $D_n^{k_i}$, where

$$D_n^k \stackrel{\text{def}}{=} E_{\mu_k} \log \frac{\mu_k(Z_0|Z_{-n+1}^{-1})}{V(Z_0|Z_{-1})} = \sum_{z_{-n+1}^0} \mu_k(z_{-n+1}^0) \log \frac{\mu_k(z_0|z_{-n+1}^{-1})}{V(z_0|z_{-1})}$$

By (4.7), $\mu_{k_i}(z_{-n+1}^0) \xrightarrow{i} \tilde{\mu}(z_{-n+1}^0)$ for all $z_{-n+1}^0 \in \mathcal{Z}^n$, and thus

$$\lim_i D_n^{k_i} = \tilde{D}_n . \quad (4.9)$$

Step 3. We observe that if $k \geq n$, then D_n^k and $D_k = D_k^k$ are related via the inequality

$$D^k - D_n^k = H_{\mu_k}(Z_0|Z_{-n+1}^{-1}) - H_{\mu_k}(Z_0|Z_{-k+1}^{-1}) \geq 0 . \quad (4.10)$$

Step 4. Combining (4.9) and (4.10) yields

$$(\forall n \geq 0) \quad \lim_i D^{k_i} \geq \lim_i D_n^{k_i} = \tilde{D}_n .$$

Taking the limit as $n \rightarrow \infty$ and using (4.8), we obtain

$$\lim_i D^{k_i} \geq \tilde{D} . \quad \triangle$$

Remarks. (a) Using the nesting property of the subspaces $\{\mathcal{L}_k\}$ it is fairly straightforward to show that the sequence $\{D^k\}$ is nondecreasing. It is also possible to construct examples in which it is strictly increasing. This implies that the bound on the error exponent given in Theorem 4.1 does not, in general, have a finite-dimensional representation.

(b) If we modify the measure $\tilde{\mu}$ to be one that almost achieves the infimum over \mathcal{L} in inequality (4.6), we can employ the monotonicity of \tilde{D}_n as established in Step 3 above and the definition of D_k to obtain the reverse inequality, i.e.,

$$\lim_k D^k \leq \inf_{\tilde{P} \in \mathcal{L}} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} .$$

This will also be deduced from the converse theorem of the following section.

5. Converse Theorem

We now prove that the two-sided single-bit compression scheme employed in the proof of Theorem 4.1 is asymptotically optimal amongst the broader class of schemes involving one-sided fixed-bit compression. In particular, the bound of Theorem 4.1 is the universal error exponent for that class.

Theorem 5.1. *If $M_X \geq 2$ and $\epsilon \in (0, 1)$, then*

$$\theta(M_X, \infty, \epsilon) \leq \inf_{\tilde{P} \in \mathcal{L}} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} .$$

Proof. Consider an arbitrary acceptance region defined by

$$\mathcal{A}_n = \bigcup_{i=1}^{M_X} C_n^{(i)} \times F_n^{(i)} ,$$

where $C_n^{(i)} \subset \mathcal{X}^n$ and $F_n^{(i)} \subset \mathcal{Y}^n$. Since $P(\mathcal{A}_n) \geq 1 - \epsilon$, there exists $j \in \{1, \dots, M_X\}$ such that

$$P(C_n^{(j)} \times F_n^{(j)}) \geq \frac{1 - \epsilon}{M_X} .$$

We write for brevity $C_n = C_n^{(j)}$, $F_n = F_n^{(j)}$ and $B_n = C_n \times F_n$, and note that for $0 \leq \lambda \leq (1 - \epsilon)/M_X$ we have

$$P(B_n) \geq \lambda .$$

We now use Lemma 3.2 to obtain a Hamming neighborhood $\Gamma^{k_n} B_n$ of B_n of probability close to unity under every measure in the class \mathcal{L} . The strong mixing condition required by the hypothesis of the lemma is satisfied by P , since the transition matrix W is irreducible and aperiodic. Thus for all $\{k_n\}$ and sufficiently large n , we have

$$P(\Gamma^{k_n} B_n) \geq 1 - \frac{n\lambda}{k_n} .$$

In what follows we use $k_n = \lceil n\sqrt{\lambda} \rceil$ and drop the subscript n from Γ^{k_n} . The previous inequality then becomes

$$P(\Gamma^k B_n) \geq 1 - \sqrt{\lambda} .$$

This readily implies

$$P(\Gamma^k C_n \times \Gamma^k F_n) \geq 1 - \sqrt{\lambda} ,$$

which in turn yields

$$P(\Gamma^k C_n) \geq 1 - \sqrt{\lambda} \quad \text{and} \quad P(\Gamma^k F_n) \geq 1 - \sqrt{\lambda} .$$

The above two inequalities also hold for any \tilde{P} in \mathcal{L} replacing P , since any such measure will agree with P on $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{Y}^{\mathbb{Z}}$. We thus have

$$\tilde{P}(\Gamma^k C_n \times \Gamma^k F_n) \geq P(\Gamma^k C_n) + P(\Gamma^k F_n) - 1 \geq 1 - 2\sqrt{\lambda} ,$$

and since $\Gamma^k C_n \times \Gamma^k F_n \subset \Gamma^{2k}(C_n \times F_n) = \Gamma^{2k} B_n$, we obtain

$$\tilde{P}(\Gamma^{2k} B_n) \geq 1 - 2\sqrt{\lambda} .$$

Next we estimate $Q(\Gamma^{2k} B_n)$ using the above bound on the \tilde{P} -probability of the same set. We have

$$\begin{aligned} Q(\Gamma^{2k} B_n) &= \sum_{z_1^n \in \Gamma^{2k} B_n} Q(z_1^n) \\ &= \sum_{z_1^n \in \Gamma^{2k} B_n} \exp[-ni_n(z_1^n)] \tilde{P}(z_1^n) , \end{aligned} \tag{5.1}$$

where

$$i_n(z_1^n) \stackrel{\text{def}}{=} \frac{1}{n} \log \frac{\tilde{P}(z_1^n)}{Q(z_1^n)} .$$

If \tilde{P} is an ergodic measure, then by virtue of the Shannon-McMillan-Breiman theorem, the sequence of random variables on $\mathcal{Z}^{\mathbf{Z}}$ induced by the mappings $\{i_n\}$ will converge almost surely to the constant D_∞ , where

$$D_\infty \stackrel{\text{def}}{=} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} .$$

It then follows easily from (5.1) that given any $\zeta > 0$, for all sufficiently large n ,

$$Q(\Gamma^{2k} B_n) \geq \exp[-n(D_\infty + \zeta)] .$$

In the general case where \tilde{P} is stationary but not ergodic, we proceed as follows. By a generalized Shannon-McMillan-Breiman theorem, the random sequence $\{i_n\}$ converges almost surely and in L^1 to a random variable i_∞ such that $E_{\tilde{P}} i_\infty = D_\infty$. Using Jensen's inequality, we obtain from (5.1)

$$Q(\Gamma^{2k} B_n) \geq \tilde{P}(\Gamma^{2k} B_n) \exp \left[-n \frac{\alpha_n}{\tilde{P}(\Gamma^{2k} B_n)} \right] ,$$

where $\alpha_n \stackrel{\text{def}}{=} E_{\tilde{P}}[i_n 1_{\Gamma^{2k} B_n}]$. We then use the L^1 convergence of i_n to i_∞ and an absolute continuity argument to conclude that if $\zeta > 0$ and n is sufficiently large, then

$$|\alpha_n - D_\infty| \leq \xi(\lambda) + \zeta$$

and

$$Q(\Gamma^{2k} B_n) \geq \exp[-n(D_\infty + \xi(\lambda) + \zeta)] , \quad (5.2)$$

where $\xi(\cdot)$ is independent of ζ, n and is such that $\xi(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$.

As a final step, we “reduce” $\Gamma^{2k} B_n$ to the original set B_n and derive a lower bound on $Q(B_n)$ with the aid of (5.2). The ratio $Q(z_1^n)/Q(\bar{z}_1^n)$ for $z_1^n \in B_n$ and $\bar{z}_1^n \in \Gamma^{2k} B_n$ is at least ρ^{4k} , where

$$\rho \stackrel{\text{def}}{=} \min_{(z_{-1}, z_0) \in \mathcal{Z}^2} V(z_0 | z_{-1}) \wedge \min_{z \in \mathcal{Z}} \pi_V(z)$$

and $\rho > 0$ by hypothesis. A standard upper bound [2] on $|\Gamma^{2k}\{z_1^n\}|$ yields

$$Q(B_n) \geq \exp[-n\nu(k_n/n)] Q(\Gamma^{2k} B_n) , \quad (5.3)$$

where $\nu(u) = h(2u) + 2u \log(|\mathcal{Z}|/\rho^2)$. The function $\nu(u)$ is concave, monotonically increasing in a neighborhood of the origin and such that $\nu(u) \rightarrow 0$ as $u \rightarrow 0$. Recalling that $k_n = \lceil n\sqrt{\lambda} \rceil$, we obtain for sufficiently small λ and sufficiently large n ,

$$\nu(k_n/n) \leq \nu(\sqrt{\lambda}) + \nu(1/n) .$$

This, combined with (5.2) and (5.3), yields for any $\zeta > 0$

$$Q(B_n) \geq \exp[-n(D_\infty + \xi(\lambda) + \nu(\sqrt{\lambda}) + \nu(1/n) + \zeta)] ,$$

and as λ can be taken arbitrarily close to 0,

$$-\limsup_n \frac{1}{n} \log Q(\mathcal{A}_n) \leq -\limsup_n \frac{1}{n} \log Q(B_n) \leq D_\infty .$$

Thus $\theta(M_X, \infty, \epsilon) \leq D_\infty$, and letting the measure \tilde{P} range over \mathcal{L} , we obtain

$$\theta(M_X, \infty, \epsilon) \leq \inf_{\tilde{P} \in \mathcal{L}} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-\infty}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} . \quad \triangle$$

Theorems 4.1 and 5.1 together yield Theorem 2.1, i.e., the complete characterization of the error exponent for all codebook sizes M_X, M_Y (provided one of the two is finite) and $\epsilon \in (0, 1)$.

6. Concluding remarks

It should be noted that the constructive argument in the proof of Theorem 4.1 establishes an asymptotically optimal sequence of acceptance regions $\{\mathcal{A}_n\}$ that is also universal, i.e. good for every level $\epsilon \in (0, 1)$, compression constraint (M_X, M_Y) and alternative distribution $Q = \pi_V \diamond V$ (provided $V > 0$). If the irreducibility assumption on W is relaxed to allow non-ergodic sources under the null hypothesis, then the error exponent and the associated regions will no longer be independent of ϵ except in trivial cases.

Our results can be easily extended to higher-order Markov processes under similar assumptions, most notably the nonnegativity constraint on the alternative distribution which is essential to our argument. An issue that remains unresolved is whether the same error exponent prevails under asymptotically zero-rate data compression, i.e., in schemes where the number of available messages is allowed to grow subexponentially with n . This is indeed the case for memoryless sources [4], but the known proof of this fact invokes the standard version of the blowing-up lemma which does not—to our knowledge—possess a counterpart for ergodic Markov sources. Thus at the present time can only employ the weaker result of Lemma 3.2, which limits the scope of our conclusions to message sets of fixed size.

REFERENCES

- [1] V. Anantharam, “A large deviations approach to error exponents in source coding and hypothesis testing,” *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 938-943, July 1990.
- [2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1982 and Budapest, Hungary: Akadémiai Kiadó, 1981.

- [3] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 759-772, Nov. 1987.
- [4] H. M. H. Shalaby and A. Papamarcou, "A converse theorem for a class of multiterminal detection problems," presented at the 1990 IEEE International Symposium on Information Theory, San Diego, California.
- [5] 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.
- [6] D. Pollard, "Convergence of Stochastic Processes," Springer, New York, 1984.
- [7] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Ann. Prob.*, vol. 16, pp. 899-909, 1988.
- [8] D. J. Rudolph, "If a two-point extension of a Bernoulli shift has an ergodic square, then it is Bernoulli," *Israel Jour. Math.*, vol. 30, pp. 159-180, 1978.
- [9] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 445-446, May 1986.