Error Exponents for Distributed Detection of Markov Sources

Hossam M. H. Shalaby, *Member, IEEE,* and Adrian Papamarcou, *Member, IEEE*

***Abstract*— The paper considers a binary hypothesis testing system in which two sensors simultaneously observe a discrete-time finite-valued stationary ergodic Markov source and transmit M-ary messages to a Neyman–Pearson central detector. The size M of the message alphabet increases at most subexponentially with the number of observations. The asymptotic behavior of the type II error rate is investigated as the number of observations increases to infinity, and the associated error exponent is obtained under mild assumptions on the source distributions. This exponent is independent of the test level ϵ and the actual codebook sizes M, is achieved by a universally optimal sequence of acceptance regions, and is characterized by an infimum of informational divergence rate over a class of infinite-dimensional distributions. Important differences—due to the observations being Markov—between the asymptotically optimal distributed tests and their nondistributed counterparts are highlighted. The converse results require a blowing-up lemma for stationary ergodic Markov sources, which is also proven.**

***Index Terms*— hypothesis testing, distributed detection, error exponent, Markov source, divergence rate, blowing-up lemma.**

I. Introduction
Problem Statement and Background

In this paper, we discuss the asymptotically optimal design of a distributed hypothesis testing system for Markov sources. Our setup is as follows:

i) a discrete-time, finite-alphabet, stationary ergodic Markov source $(X_i, Y_i)_{-\infty}^{\infty}$ with $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$;

ii) two remote sensors S_X and S_Y;

iii) a central detector.

The sensors S_X and S_Y observe the source components X_1^n and Y_1^n, and encode their observations into single messages taking $M_{X,n}$ and $M_{Y,n}$ values, respectively. 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. Our aim is to study the asymptotic performance of the optimal test of level $\epsilon \in (0, 1)$ 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

$$\theta(M_X, M_Y, \epsilon) \stackrel{\text{def}}{=} -\limsup_n \frac{1}{n}\log \beta_n(M_X, M_Y, \epsilon). \quad (1.1)$$

The codebook sizes $M_{X,n}$ and $M_{Y,n}$ are essential parameters in the above formulation and determine the extent to which the detection process is distributed. For example, setting $M_{X,n} = \mathcal{X}^n$ and $M_{Y,n} = \mathcal{Y}^n$ we obtain the case of the conventional centralized detector, for which the error exponent is known [1]. Specifically, if the source has transition matrix $W(\cdot \mid \cdot)$ under H_0 and $V(\cdot \mid \cdot)$ under H_1, then

$$\theta(M_X, M_Y, \epsilon) = D(W \,\|\, V) \quad (1.2)$$

where $D(W \,\|\, V)$ is the *conditional informational divergence* [2] of $W(\cdot \mid \cdot)$ to $V(\cdot \mid \cdot)$. It is defined by

$$D(W \,\|\, V) = \sum_{(z_1, z_2) \in \mathcal{Z}^2} \pi(z_1) W(z_2 \mid z_1) \log \frac{W(z_2 \mid z_1)}{V(z_2 \mid z_1)} \quad (1.3)$$

where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\pi(\cdot)$ is the initial distribution of the source under H_0. If we assume that $W(\cdot \mid \cdot)$ is irreducible, then $\pi(\cdot)$ is uniquely determined by W, and the notation $D(W \,\|\, V)$ is unambiguous. We note from (1.2) and (1.3) that the value of the error exponent does not involve ϵ and depends on the source distribution only through its restriction to two consecutive time coordinates.

In this paper we consider the nondegenerate case of one-sided or two-sided data compression at *asymptotically zero rate*. This means that one or both codebooks grow at most subexponentially with n:

$$\lim_n \frac{1}{n}\log M_{X,n} = 0 \quad \text{and/or} \quad \lim_n \frac{1}{n}\log M_{Y,n} = 0. \quad (1.4)$$

Our search for $\theta(M_X, M_Y, \epsilon)$ follows earlier work [3], [4] on memoryless sources exhibiting spatial dependence. These studies showed that if $(X_i, Y_i)_{-\infty}^{\infty}$ is an i.i.d. process whose distribution is 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

$$\theta_{\text{iid}}(M_X, M_Y, \epsilon) = \min_{\tilde{P}_{XY}:\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y} D(\tilde{P}_{XY} \,\|\, Q_{XY}) \quad (1.5)$$

provided (1.4) holds. Here $D(\cdot\|\cdot)$ is the ordinary (not conditional) *informational divergence*, defined for two distributions

Manuscript received December 16, 1991; revised October 12, 1992. This work was supported by the Institute of Systems Research (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 25th Annual Conference on Information Sciences and Systems, held March 20–22, 1991 at the Johns Hopkins University, Baltimore, MD, USA.

H. M. H. Shalaby was with the Electrical Engineering Department and the Institute for Systems Research, University of Maryland, College Park, MD 20742. He is now with the Department of Electrical Engineering, Faculty of Engineering, University of Alexandria, Alexandria 21544, Egypt.

A. Papamarcou is with the Electrical Engineering Department, University of Maryland, College Park, MD 20742 USA.

IEEE Log Number 9400235.

P and Q on $\mathcal{Z}$ by

$$D(P\,\|\,Q) = \sum_{z\in\mathcal{Z}} P(z)\log\frac{P(z)}{Q(z)}. \tag{1.6}$$

The value of the error exponent does not involve $\{M_{X,n}\}$, $\{M_{Y,n}\}$ or ϵ, and depends on the source distribution only through its restriction to a single time coordinate.

II. Main Contribution

Based on the results presented in the previous section, one might conjecture that for distributed testing of Markov hypotheses under the zero rate constraint (1.4):

i) the error exponent $\theta(M_X, M_Y, \epsilon)$ is independent of $\{M_{X,n}\}$, $\{M_{Y,n}\}$ and ϵ;

ii) a characterization of $\theta(M_X, M_Y, \epsilon)$ can be given via the minimum of a suitable divergence functional over a 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 $V(\cdot\mid\cdot)$ satisfies the positivity constraint $V(\cdot\mid\cdot) > 0$ and condition (1.4) holds, then the error exponent $\theta(M_X, M_Y, \epsilon)$ is given by the infimum of

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

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

The expectation appearing in the statement of the theorem equals the *divergence rate* of the stationary measure $\tilde{P}$ relative to a stationary Markov measure with transition matrix $V(\cdot\mid\cdot)$. The above characterization of $\theta(M_X, M_Y, \epsilon)$ in terms of this divergence rate clearly involves a class of distributions on an infinite-dimensional space and does not appear to have a finite-dimensional equivalent. Thus statement ii) seems to be false.

This conclusion is not implausible considering that the source components $(X_i)_{-\infty}^{\infty}$ and $(Y_i)_{-\infty}^{\infty}$ are *not*, in general, Markov of any finite-order (even though the joint process $(X_i, Y_i)_{-\infty}^{\infty}$ is). Yet as it turns out, non-Markovity of components is not critical here; the error exponent also seems to have an irreducibly infinite-dimensional characterization for a large class of examples in which the X and Y processes *are* individually Markov. This implies that in an asymptotically optimal system, the sensors cannot rely on the empirical transition matrix alone in order to encode their Markov observations. In other words, the usual sufficient statistic for the detection of Markov sources in the conventional framework does not necessarily yield optimal results in a distributed one.

The paper is organized as follows. Section III covers technical preliminaries. The direct part of Theorem 2.1 is established in Section IV, and the converse part in Section V. Sections VI and VII contains a discussion of our results and their immediate extensions. The class of examples mentioned in the previous paragraph is developed in Appendix A. Proof of an auxiliary result (the blowing-up lemma for stationary ergodic Markov sources) is given in Appendix B.

III. Preliminaries

A. General Notation

For simplicity we let $\mathcal{Z}\overset{\text{def}}{=}\mathcal{X}\times\mathcal{Y}$ and $Z_i\overset{\text{def}}{=}(X_i, Y_i)$. We will denote $(Z_i,\cdots,Z_j)$ and $(z_i,\cdots,z_j)$ by Z_i^j and z_i^j, respectively.

If P is a stationary measure on the Borel field of $\mathcal{Z}^{\boldsymbol{Z}}$, we will use P_n to denote the restriction of P to n successive time coordinates. Where $A_n\subset\mathcal{Z}^n$, we will write for simplicity $P(A_n)$ instead of $P_n(A_n)$. Also, if $A\subset\mathcal{X}^n$, we will write $P(A)$ instead of $P_n(A\times\mathcal{Y}^n)$; and similarly for $B\subset\mathcal{Y}^n$.

The measure P can be specified by the family of conditional distributions $\{W_n,\ n\in\boldsymbol{N}\}$ defined by

$$W_n(z_n\mid z_1^{n-1}) = \frac{P(z_1^n)}{P(z_1^{n-1})}$$

where, as usual, $W_n(\cdot\mid z_1^{n-1})$ is chosen arbitrarily when $P(z_1^{n-1}) = 0$. We thus have

$$P(z_1^n) = \prod_{i=1}^{n} W_i(z_i\mid z_1^{i-1})$$

which leads to the abbreviated representation $P = \prod_i W_i$.

If P is first-order Markov, then W_1 (the initial distribution) and W_2 (the transition matrix) completely specify P; while W_2 alone is sufficient if it is irreducible. Similarly, if P is order-$(k-1)$ Markov, then it is completely specified by $W_1,\cdots,W_k$; and by W_k alone if W_k is irreducible.

B. Entropy, Divergence, and Ergodic Decomposition

We have the following identities for entropy, conditional entropy, and entropy rate (respectively) pertaining to the stationary measure $P = \prod_i W_i$ on $\mathcal{Z}^{\boldsymbol{Z}}$:

$$\begin{aligned} H(P_n) &=: -E_P\log P(Z_1^n) = -\sum_{z_1^n\in\mathcal{Z}^n} P(z_1^n)\log P(z_1^n)\\ H(W_n) &=: -E_P\log W_n(Z_n\mid Z_1^{n-1})\\ &=: -\sum_{z_1^n\in\mathcal{Z}^n} P(z_1^n)\log W(z_n\mid z_1^{n-1})\\ \overline{H}(P) &=: \lim_n\frac{1}{n}H(P_n) = \lim_n H(W_n). \end{aligned} \tag{3.1}$$

If P is order-$(k-1)$ Markov, then $H(W_n) = H(W_k)$ for $n\geq k$ and thus $\overline{H}(P) = H(W_k)$.

Now let $Q = \prod_i V_i$ be another stationary measure on $\mathcal{Z}^{\boldsymbol{Z}}$. The divergence of P_n relative to Q_n is defined by

$$D(P_n\,\|\,Q_n) = E_P\log\frac{P(Z_1^n)}{Q(Z_1^n)} = \sum_{z_1^n\in\mathcal{Z}^n} P(z_1^n)\log\frac{P(z_1^n)}{Q(z_1^n)}$$

and the *conditional divergence* of W_n relative to V_n by

$$D(W_n \| V_n) = E_P \log \frac{W(Z_n \mid Z_1^{n-1})}{V(Z_n \mid Z_1^{n-1})} = \sum_{z_1^n \in \mathcal{Z}^n} P(z_1^n) \log \frac{W(z_n \mid z_1^{n-1})}{V(z_n \mid z_1^{n-1})}.$$

If Q_n is finite-order Markov and $Q_n \gg P_n$ for all n, then the *divergence rate* of P relative to Q also exists and is given by (see also Lemma 7.4.1 in [5])

$$\overline{D}(P \| Q) = \lim_n \frac{1}{n} D(P_n \| Q_n) = \lim_n D(W_n \| V_n). \quad (3.2)$$

If, in particular, both P and Q are order-$(k-1)$ Markov, then $D(P_n \| Q_n) = D(W_k \| V_k)$ for $n \geq k$ and thus $\overline{D}(P \| Q) = D(W_k \| V_k)$.

The following facts on ergodic decompositions can be found in [5]. There exists a family of stationary ergodic measures $\{m_z,\ z \in \mathcal{Z}^{\boldsymbol{Z}}\}$ on $\mathcal{Z}^{\boldsymbol{Z}}$ such that any stationary measure P on $\mathcal{Z}^{\boldsymbol{Z}}$ can be expressed in the form

$$P(\cdot) = \int m_z(\cdot)\, dP(z).$$

If P itself is ergodic, then $m_z = P$ with P-probability 1. The entropy rate of P has a similar decomposition as

$$\overline{H}(P) = \int \overline{H}(m_z)\, dP(z).$$

If Q is finite-order Markov and $Q_n \gg P_n$ for all n, then with P-probability 1 the measure m_z satisfies $Q_n \gg m_{z,n}$ for all n. The divergence rate of P relative to Q can then be written as

$$\overline{D}(P \| Q) = \int \overline{D}(m_z \| Q)\, dP(z).$$

C. Typical Sequences

We give here a summary of pertinent facts on a Markov concept of typicality derived from the work of Davisson, Longo and Sgarro [6]. A related concept has been used in [7].

The *order-k type* of a finite sequence $z_1^n \in \mathcal{Z}^n$ is the empirical distribution 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 all lower dimensional distributions of an order-k type $\hat{P}_k$ are shift-invariant. That is to say, if both I and $I+1 \stackrel{\text{def}}{=} \{i+1:\ i \in I\}$ are subsets of $\{1, \cdots, k\}$, then the marginals of $\hat{P}_k$ corresponding to the index sets I and $I+1$ are identical.

The above observation enables us to extend $\hat{P}_k$ to a stationary ergodic measure on $\mathcal{Z}^{\boldsymbol{Z}}$ in the following standard fashion. First, we let

$$\hat{P}_{k-1}(z_1^{k-1}) = \sum_{z_k \in \mathcal{Z}} \hat{P}_k(z_1^k)$$

and

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

Then we define a measure $\hat{P}$ on positive-time sequences by

$$(\forall n \geq k) \quad \hat{P}(Z_1 = z_1, \cdots, Z_n = z_n) = \hat{P}_{k-1}(z_1^{k-1}) \prod_{i=k}^{n} \hat{W}_k(z^i \mid z_{i-k+1}^{i-1}).$$

Using the sift-invariance property discussed earlier, it is straightforward to prove that $\hat{P}$ is stationary on $\mathcal{Z}^{\boldsymbol{N}}$ (it can then be extended to $\mathcal{Z}^{\boldsymbol{Z}}$). By construction, $\hat{P}$ is also order-$(k-1)$ Markov with transition probability matrix $\hat{W}_k(\cdot \mid \cdot)$. And since $\hat{P}_k$ is obtained by evaluating the relative frequency of k-strings along a period sequence, the measure $\hat{P}$ has a single recurrent class of $(k-1)$-strings; it is thus ergodic.

By the ergodic theorem, for any stationary ergodic measure $\tilde{P}$ on $\mathcal{Z}^{\boldsymbol{Z}}$, there exists a suitably long sequence z_1^n whose order-k type approximates $\tilde{P}_k$ (e.g., in sup-norm) arbitrarily closely. Thus, the set $\mathcal{P}_k(\mathcal{Z}^n)$ of all order-k types obtained from sequences of length n is asymptotically dense in the set of all measures on $\mathcal{Z}^k$ that have stationary ergodic extensions on $\mathcal{Z}^{\boldsymbol{Z}}$.

We will distinguish between two notions of *typicality*. The first is precise: if $\hat{P}_k \in \mathcal{P}_k(\mathcal{Z}^n)$, we say that a sequence z_1^n is $\hat{P}_k$-typical if its order-k type equals $\hat{P}_k$. We denote the set of all such sequences by $\hat{T}_k(\mathcal{Z}^n)$ or, where no confusion as to the value of n may arise, simply by $\hat{T}_{Z,k}$. The second notion of typicality involves an approximation: if $\tilde{P}$ is an arbitrary stationary distribution on $\mathcal{Z}^{\boldsymbol{Z}}$, we say that a sequence z_1^n is $(\tilde{P}_k, \eta)$-typical if its type $\hat{P}_k$ satisfies

$$\max_{a_1^k} \left| \hat{P}_k(a_1^k) - \tilde{P}(a_1^k) \right| \leq \eta$$

and we denote the set of all such sequences by $\tilde{T}_{k,\eta}(\mathcal{Z}^n)$, or simply by $\tilde{T}_{Z,k,\eta}$.

In the following lemma we give some standard facts on the cardinalities (denoted by $|\cdot|$) and probabilities of some of the sets introduced previously. The proof of assertions i) and ii) for $k=2$ can be found in [6]; generalization to arbitrary k is straightforward. Assertion iii) is easily established using the pointwise ergodic theorem.

Lemma 3.1:

i) $|\mathcal{P}_k(\mathcal{Z}^n)| \leq r(n)$ where $r(\cdot)$ is a polynomial of degree $|\mathcal{Z}|^k$.

ii) Let $\hat{P} = \prod_i \hat{W}_i$ be a stationary measure such that $\hat{P}_k \in \mathcal{P}_k(\mathcal{Z}^n)$. Then there exists a polynomial $s(\cdot)$ of degree $2|\mathcal{Z}|^k$ and an absolute constant c such that

$$\frac{1}{s(n)} \exp[nH(\hat{W}_k)] \leq |\hat{T}_{Z,k}| \leq c \exp[nH(\hat{W}_k)].$$

iii) Let $\hat{P}$ be a stationary ergodic measure on $\mathcal{Z}^{\boldsymbol{Z}}$. Then for any integer $k \geq 1$ there exist positive sequences $\{\eta_n\}_{n=1}^{\infty}$

and $\{\xi_n\}_{n=1}^{\infty}$ converging to zero such that the set of $(\tilde{P}_k, \eta_n)$-typical sequences of length n satisfies

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

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]. This lemma has been recently extended by Marton and Shields [8], [9] to all stationary ergodic sources which are finitary codings of i.i.d. sources. We employ the following weaker version of their result, for which we give independent proof in Appendix B.

Lemma 3.2: Let P be a stationary first-order Markov measure on $\mathcal{Z}^{\mathbf{Z}}$ with irreducible and aperiodic transition matrix. If, for $\delta_n \to 0$, the set $B_n \subset \mathcal{Z}^n$ satisfies

$$(\forall n) \qquad P(B_n) \geq \exp[-n\delta_n],$$

then there exist integers κ_n with $\kappa_n/n \to 0$ such that the Hamming κ_n-neighborhood of B_n—denoted by $\Gamma^{\kappa_n} B_n$—satisfies

$$\lim_n P(\Gamma^{\kappa_n} B_n) = 1.$$

IV. Direct Theorem

We recapitulate the problem statement as follows. We are given a stationary ergodic first-order Markov source $Z_{-\infty}^{\infty} = (X_i, Y_i)_{-\infty}^{\infty}$ and two simple hypotheses H_0 and H_1. Under H_0, the source distribution is $P = \prod_i W_i$ where $W = W$ is an irreducible aperiodic transition matrix. Under H_1, we have $Q = \prod_i V_i$ where $V_2 = V$ satisfies the additional positivity constraint $V > 0$. The condition $V > 0$, which will be needed in the proof of the converse theorem, also guarantees that $Q_n \gg \tilde{P}_n$ for any stationary measure $\tilde{P}$ and value of n.

The two sensors S_X and S_Y encode their observations X_1^n and Y_1^n into one of $M_{X,n}$ and $M_{Y,n}$ messages, respectively, where $M_{X,n}$ and $M_{Y,n}$ satisfy the asymptotic zero rate constraint (1.4). Thus, for any given n, S_X partitions the space $\mathcal{X}^n$ into cells $C_n^{(i)}$ where $1 \leq i \leq M_{X,n}$; and S_Y partitions $\mathcal{Y}^n$ into cells $G_n^{(j)}$, where $1 \leq j \leq M_{Y,n}$. Each sensor then communicates to the central detector the cell index corresponding to its observation. This forces the central detector to employ an acceptance region (for the null hypothesis) of the form

$$\mathcal{A}_n = \bigcup_{i=1}^{M_{X,n}} C_n^{(i)} \times F_n^{(i)} \tag{4.1}$$

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

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
- are expressible in the form (4.1) using partitions $\{C_n^{(i)}\}$ and $\{G_n^{(j)}\}$ constrained in size by (1.4).

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 (1.1).

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 $\mathcal{Z}^{\mathbf{Z}}$. These are defined by

$$\mathcal{L}_k = \{\tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbf{Z}}: \tilde{P}_k \text{ and } P_k \text{ agree on } \mathcal{X}^k \text{ and } \mathcal{Y}^k\}; \tag{4.2}$$

$$\mathcal{L} = \{\tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbf{Z}}: \tilde{P} \text{ and } P \text{ agree on } \mathcal{X}^{\mathbf{Z}} \text{ and } \mathcal{Y}^{\mathbf{Z}}\}. \tag{4.3}$$

It is clear that $\mathcal{L}_k \supset \mathcal{L}_{k+1}$ and $\mathcal{L}_k \searrow \mathcal{L}$. Using the notation developed in the previous section, we will express the generic element of the above subspaces as $\tilde{P} = \prod_i \tilde{W}_i$.

Theorem 4.1: If we let

$$D_k \stackrel{\text{def}}{=} \min_{\tilde{P} \in \mathcal{L}^k} D(\tilde{W}_k \,\|\, V_k)$$

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

$$\theta(2, 2, \epsilon) \geq \sup_k D_k \geq \inf_{\tilde{P} \in \mathcal{L}} \overline{D}(\tilde{P} \,\|\, Q).$$

Proof: The idea is to construct for fixed k a sequence of acceptance regions $\mathcal{A}_n \subset \mathcal{Z}^n$ that contain the set $T_{Z,k,\eta}$ of (P_k, η)-typical sequences where $\eta = \eta_n \to 0$ is as in statement iii) of Lemma 3.1. This will ensure that $P(\mathcal{A}_n)$ is greater than $1 - \epsilon$ for all sufficiently large n, and thus the type I error constraint will be satisfied.

The set $T_{Z,k,\eta}$ itself cannot be expressed in the form of (4.1) if $M_{X,n}$ and $M_{Y,n}$ satisfy constraint (1.4), and thus the choice $\mathcal{A}_n = T_{Z,k,\eta}$ is not permissible. We consider instead the restriction of P on $\mathcal{X}^{\mathbf{Z}}$ and $\mathcal{Y}^{\mathbf{Z}}$, and let

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

Here $T_{X,k,\eta}$ and $T_{Y,k,\eta}$ are the sets of (P_k, η)-typical sequences in $\mathcal{X}^n$ and (P_k, η)-typical sequences in $\mathcal{Y}^n$, respectively. Clearly $\mathcal{A}_n$ can be written as in (4.1) with $M_{X,n} = M_{Y,n} = 2$, and thus it satisfies constraint (1.4).

The type I error constraint is met for n sufficiently large, since $\mathcal{A}_n \supset T_{Z,k,\zeta}$ for ζ_n equal to a suitable multiple of η_n. To evaluate the corresponding type II error, we note z_1^n lies in $\mathcal{A}_n$ if and only if its order-k type is drawn from the class

$$\Phi_k \stackrel{\text{def}}{=} \left\{ \hat{P}_k \in \mathcal{P}_k(\mathcal{Z}^n): \max_{x_1^k} \left| \hat{P}_k(x_1^k) - P(x_1^k) \right| \leq \eta_n, \max_{y_1^k} \left| \hat{P}_k(y_1^k) - P(y_1^k) \right| \leq \eta_n \right\} \tag{4.4}$$

and thus

$$\mathcal{A}_n = \bigcup_{\hat{P}_k \in \Phi_k} \hat{T}_{Z,k}. \tag{4.5}$$

A routine computation based on Lemma 3.1 ii) gives for any $\hat{P}_k \in \mathcal{P}_k(\mathcal{Z}^n)$

$$\frac{1}{\sigma(n)}\exp[-nD(\hat{W}_k \| V_k)] \le Q(\hat{T}_{Z,k}) \le \gamma \exp[-nD(\hat{W}_k \| V_k)] \quad (4.6)$$

where $\sigma(n)$ is a polynomial of degree at most $2|\mathcal{Z}|^k$ and γ is an absolute constant. From (4.4), (4.5), (4.6) and the type counting result in Lemma 3.1 i) we obtain, for $\delta_n \to 0$, the estimate

$$\frac{1}{n}\log Q(\mathcal{A}_n) \le -\min_{\hat{P}_k \in \Phi_k} D(\hat{W}_k \| V_k) + \delta_n.$$

Consider now the classes of measures on $\mathcal{Z}^k$ given by

$$\mathcal{S}_k = \{\tilde{P}_k\colon\ \tilde{P} \in \mathcal{L}_k\} \quad \text{and} \quad \mathcal{E}_k = \{\tilde{P}_k\colon\ \tilde{P} \text{ ergodic in } \mathcal{L}_k\}.$$

The class $\mathcal{E}_k$ is nonempty (it contains $\tilde{P} = P_X \times P_Y$ where P_X and P_Y are the ergodic X and Y marginals of the null measure P) and the same is true of $\mathcal{S}_k \supset \mathcal{E}_k$. It is easy to verify that S_k is closed (e.g., in sup-norm) and that any measure in $\mathcal{S}_k$, when contaminated by $(P_X \times P_Y)_k$, yields a measure in $\mathcal{E}_k$; thus $\mathcal{S}_k$ is the closure of $\mathcal{E}_k$. From the discussion above and in Section III-C, the classes Φ_k and $\mathcal{S}_k$ will approximate each other as $n \to \infty$ and $\eta_n \to 0$, in that the distance between any measure in one class and the closest measure in the other class will approach zero. The continuity of the conditional divergence functional then gives

$$\frac{1}{n}\log Q(\mathcal{A}_n) \le -\min_{\tilde{P} \in \mathcal{L}_k} D(\tilde{W}_k \| V_k) + \delta_n + \nu(\eta_n)$$

where $\nu(\eta_n) \to 0$ as $n \to \infty$. Invoking the definition of the error exponent, we obtain

$$\begin{aligned}\theta(2, 2, \epsilon) &= -\limsup_n \frac{1}{n}\log \beta_n(2, 2, \epsilon) \\ &\ge -\limsup_n \frac{1}{n}\log Q(\mathcal{A}_n) \\ &\ge \min_{\tilde{P} \in \mathcal{L}_k} D(\tilde{W}_k \| V_k) = D_k.\end{aligned}$$

It also follows that $\theta(2, 2, \epsilon) \ge \sup_k D_k$. This establishes the first inequality in the statement of the theorem. (For the above choice of $\mathcal{A}_n$, it is straightforward to show that $-\lim_n n^{-1}\log Q(\mathcal{A}_n)$ also exists and is equal to D_k.)

Note that $\mathcal{L}_k \supset \mathcal{L}_{k+1}$ implies $D_{k+1} \ge D_k$, so that $\sup_k D_k = \lim_k D_k$. Thus, it remains to show that

$$\lim_k D_k \ge \inf_{\tilde{P} \in \mathcal{L}} \overline{D}(\tilde{P} \| Q). \quad (4.7)$$

To do so, we consider a sequence of measures $\{\mu^{(k)}\}$ on $\mathcal{Z}^{\boldsymbol{Z}}$ such that $\mu^{(k)} = \prod_i \omega_i^{(k)}$ achieves the minimum in the definition of D_k, i.e.,

$$D_k = D(\omega_k^{(k)} \| V_k).$$

The product space $\mathcal{Z}^{\boldsymbol{Z}}$ is compact under the product topology induced by the discrete topology on $\mathcal{Z}$. It is also clearly metrizable and thus equivalent to a compact metric space. We invoke Prohorov's theorem [10] to conclude that $\{\mu^{(k)}\}$ contains a subsequence $\{\mu^{(k_i)},\ i \in \boldsymbol{N}\}$ which converges weakly to a measure μ. Weak convergence implies that every cylinder set $K \subset \mathcal{Z}^{\boldsymbol{Z}}$—being both open and closed in the product topology—will satisfy

$$\lim_i \mu^{(k_i)}(K) = \mu(K). \quad (4.8)$$

It also follows easily that μ is stationary and lies in $\mathcal{L}$. We will write $\mu = \prod_i \omega_i$.

To establish (4.7), it suffices to show that $\lim_i D_{k_i} \ge \overline{D}(\mu \| Q)$. We do so in four steps.

Step 1) We approximate $\overline{D}(\mu \| Q)$ by $D(\omega_n \| V_n)$. Indeed, from (3.2) we have

$$\lim_n D(\omega_n \| V_n) = \overline{D}(\mu \| Q). \quad (4.9)$$

Although the above suffices for our purposes, we also note that

$$D(\omega_{n+1} \| V_{n+1}) - D(\omega_n \| V_n) = H(\omega_n) - H(\omega_{n+1}) \ge 0$$

and hence

$$D(\omega_n \| V_n) \uparrow \overline{D}(\mu \| Q).$$

Step 2) We approximate $D(\omega_n \| V_n)$ by $D(\omega_n^{(k)} \| V_n)$. By (4.8),

$$(\forall z_1^n \in \mathcal{Z}^n) \qquad \lim_i \mu^{(k_i)}(z_1^n) = \mu(z_1^n),$$
$$(\forall z_1^n\colon\ \mu(z_1^n) > 0) \qquad \lim_i \omega_n^{(k_i)}(\cdot \mid z_1^{n-1}) = \omega_n(\cdot \mid z_1^{n-1}).$$

Thus,

$$\begin{aligned}&\lim_i D(\omega_i^{(k_i)} \| V_n) \\ &\quad = \sum_{z_1^n} \lim_i \mu^{(k_i)}(z_1^n) \log \frac{\omega_n^{(k_i)}(z_n \mid z_1^{n-1})}{V(z_n \mid z_{n-1})} \\ &\quad = \sum_{z_1^n} \mu(z_1^n) \log \frac{\omega_n(z_n \mid z_1^{n-1})}{V(z_n \mid z_{n-1})} = D(\omega_n \| V_n). \quad (4.10)\end{aligned}$$

Step 3) We observe that if $k \ge n$, then $D(\omega_n^{(k)} \| V_n)$ and $D_k = D(\omega_k^{(k)} \| V_k)$ are related via

$$D_k - D(\omega_n^{(k)} \| V_n) = H(\omega_n^{(k)}) - H(\omega_k^{(k)}) \ge 0. \quad (4.11)$$

Step 4) Combining (4.10) and (4.11) yields

$$\lim_i D_{k_i} \ge \lim_i D(\omega_n^{(k_i)} \| V_n) = D(\omega_n \| V_n).$$

Taking the limit as $n \to \infty$ and using (4.9), we obtain

$$\lim_i D_{k_i} \ge \overline{D}(\mu \| Q). \qquad \triangle$$

Remarks:

a) As is the case with the definition of $\bar{\rho}$-distance [11], the infimum in (4.7) can be taken over the subclass $\mathcal{L}_e$ of ergodic measures in $\mathcal{L}$. Indeed, let $\tilde{P}$ be any measure in $\mathcal{L}$ and $\{m_z,\ z \in \mathcal{Z}^{\mathbf{Z}}\}$ be the class of ergodic measures introduced in Section III-B. One can show by a standard argument (see, e.g., [12, Theorem 8.3.1]) that the event

$$A = \{z:\ m_z \in \mathcal{L}_e\}$$

has $\tilde{P}$-probability 1. Using the ergodic decomposition of divergence rate, we then obtain

$$\begin{aligned}\overline{D}(\tilde{P} \,\|\, Q) &= \int_A \overline{D}(m_z \,\|\, Q)\, d\tilde{P}(z) \\ &\geq \inf_{z \in A} \overline{D}(m_z \,\|\, Q) \geq \inf_{\tilde{P}' \in \mathcal{L}_e} \overline{D}(\tilde{P}' \,\|\, Q),\end{aligned}$$

as needed.

b) For an arbitrary $\tilde{P} = \prod_i \tilde{W}_i$ in $\mathcal{L} \supset \mathcal{L}_k$, the monotonicity of $D(\tilde{W}_n \,\|\, V_n)$ in n as established in Step 1) implies that

$$D_k \leq D(\tilde{W}_k \,\|\, V_k) \leq \overline{D}(\tilde{P} \,\|\, Q)$$

and thus

$$\sup_k D_k \leq \inf_{\tilde{P} \in \mathcal{L}} \overline{D}(\tilde{P} \,\|\, Q).$$

This is the reverse of the inequality (4.7), and will also follow from the converse theorem below.

V. Converse Theorem

We now prove that no scheme involving one-sided or two-sided zero-rate compression can result in an error exponent higher than the lower bound of Theorem 4.1.

Theorem 5.1: If the asymptotic zero-rate compression constraint (1.4) is satisfied, then

$$\theta(M_X, M_Y, \epsilon) \leq \inf_{\tilde{P} \in \mathcal{L}_e} \overline{D}(\tilde{P} \,\|\, Q).$$

Proof: The argument parallels the proof of Theorem 3.1 in [4]. We assume without loss of generality that $n^{-1} \log M_{X,n} \to 0$ and that $M_{Y,n}$ is unconstrained.

Consider an arbitrary acceptance region defined by

$$\mathcal{A}_n = \bigcup_{i=1}^{M_{X,n}} C_{n,i} \times F_{n,i}$$

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 $j \in \{1, \cdots, M_{X,n}\}$ such that

$$P(C_{n,j} \times F_{n,j}) \geq \frac{1-\epsilon}{M_{X,n}}.$$

We write for brevity $C_{n,j} = C_n$, $F_{n,j} = F_n$ and $B_n = C_n \times F_n$. By (1.4), there exists $\delta_n \to 0$ such that

$$P(B_n) \geq \exp[-n\delta_n].$$

Next we apply Lemma 3.2 to obtain a Hamming neighborhood of B_n with probability asymptotically approaching unity under every measure in the class $\mathcal{L}_e$. Indeed, since P is stationary Markov with irreducible aperiodic transition matrix W, there exists an integer sequence $\{\kappa_n\}$ such that

$$\frac{\kappa_n}{n} \to 0 \quad \text{and} \quad \lim_n P(\Gamma^{\kappa_n} B_n) = 1.$$

Dropping the subscript n from Γ^{κ_n}, we have

$$\lim_n P(\Gamma^{\kappa} C_n \times \Gamma^{\kappa} F_n) \geq \lim_n P(\Gamma^{\kappa} B_n) = 1,$$

which in turn yields

$$\lim_n P(\Gamma^{\kappa} C_n) = 1 \quad \text{and} \quad \lim_n P(\Gamma^{\kappa} F_n) = 1.$$

The above relationships also hold for any $\tilde{P} = \prod_i \tilde{W}_i$ in $\mathcal{L}_e$ replacing P, since any such measure agrees with P on $\mathcal{X}^{\mathbf{Z}}$ and $\mathcal{Y}^{\mathbf{Z}}$. Thus,

$$\tilde{P}(\Gamma^{\kappa} C_n \times \Gamma^{\kappa} F_n) \geq \tilde{P}(\Gamma^{\kappa} C_n) + \tilde{P}(\Gamma^{\kappa} F_n) - 1 \to 1,$$

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

$$\lim_n \tilde{P}(\Gamma^{2\kappa} B_n) = 1. \tag{5.1}$$

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

$$\begin{aligned}Q(\Gamma^{2\kappa} B_n) &= \sum_{z_1^n \in \Gamma^{2\kappa} B_n} Q(z_1^n) \\ &= \sum_{z_1^n \in \Gamma^{2\kappa} B_n} \exp[-n i_n(z_1^n)] \tilde{P}(z_1^n)\end{aligned} \tag{5.2}$$

where

$$\begin{aligned}i_n(z_1^n) &\stackrel{\text{def}}{=} \frac{1}{n} \log \frac{\tilde{P}(z_1^n)}{Q(z_1^n)} \\ &= \frac{1}{n} \log \tilde{P}(z_1^n) - \frac{1}{n} \log Q(z_1) \\ &\quad - \frac{1}{n} \sum_{j=2}^{n} \log V(z_j \mid z_{j-1}).\end{aligned}$$

Since $\tilde{P}$ is ergodic, we apply the Shannon–McMillan–Breiman and ergodic theorems to conclude that the sequence of random variables on $\mathcal{Z}^{\mathbf{Z}}$ induced by the mappings $\{i_n\}$ converges $\tilde{P}$-almost surely to the constant

$$\begin{aligned}&-\overline{H}(\tilde{W}) - E_{\tilde{P}} \log V(Z_2 \mid Z_1) \\ &\quad = \lim_n E_{\tilde{P}}[\log \tilde{W}_n(Z_n \mid Z_1^{n-1}) - \log V_n(Z_n \mid Z_1^{n-1})] \\ &\quad = \overline{D}(\tilde{P} \,\|\, Q)\end{aligned}$$

[see also (3.1) and (3.2)]. It then follows easily from (5.1) and (5.2) that

$$Q(\Gamma^{2\kappa} B_n) \geq \exp[-n(\overline{D}(\tilde{P} \,\|\, Q) + \zeta_n)] \tag{5.3}$$

where $\zeta_n \to 0$ as $n \to \infty$.

As a final step,we "reduce" $\Gamma^{2\kappa} B_n$ to the original set B_n and derive a lower bound on $Q(B_n)$ with the aid of (5.3). The ratio $Q(z_1^n)/Q(\overline{z}_1^n)$ for $z_1^n \in B_n$ and $\overline{z}_1^n \in \Gamma^{2\kappa} B_n$ is at least ρ^{4k} where

$$\rho \stackrel{\text{def}}{=} \min_{z_1^2 \in \mathcal{Z}^2} V(z_2 \mid z_1) \wedge \min_{z \in \mathcal{Z}} Q(z)$$

and $\rho > 0$ by hypothesis. A standard upper bound [2, Lemma 5.1] on the size of $\Gamma^{2\kappa}\{z_1^n\}$ then yields

$$Q(B_n) \geq \exp[-n\nu(\kappa_n/n)]Q(\Gamma^{2\kappa}B_n) \tag{5.4}$$

where $\nu(u) = h(2u) + 2u\log(|\mathcal{Z}|/\rho^2)$. As n tends to infinity, both κ_n/n and $\nu(\kappa_n/n)$ tend to zero. Equations (5.3) and (5.4) then give

$$Q(B_n) \geq \exp[-n(\overline{D}(\tilde{P}\,\|\,Q) + \xi_n)]$$

where $\xi_n \to 0$ as $n \to \infty$, and hence

$$\begin{aligned} -\limsup_n \frac{1}{n}\log Q(\mathcal{A}_n) &\leq -\limsup_n \frac{1}{n}\log Q(B_n) \\ &\leq \overline{D}(\tilde{P}\,\|\,Q). \end{aligned}$$

Thus, $\theta(M_X, M_Y, \epsilon) \leq \overline{D}(\tilde{P}\,\|\,Q)$, and also

$$\theta(M_X, M_Y, \epsilon) \leq \inf_{\tilde{P}\in\mathcal{L}_e} \overline{D}(\tilde{P}\,\|\,Q). \qquad \Delta$$

Theorems 4.1, 5.1 and the first remark following the proof of Theorem 4.1 together yield

$$\theta(M_X, M_Y, \epsilon) = \sup_k D_k = \inf_{\tilde{P}\in\mathcal{L}_e} \overline{D}(\tilde{P}\,\|\,Q).$$

From Lemma 7.4.1 in [5], we have

$$\overline{D}(\tilde{P}\,\|\,Q) = E_{\overline{P}}\log\frac{\tilde{P}\{(XY)_0 \mid (XY)_{-\infty}^{-1}\}}{V\{(XY)_0 \mid (XY)_{-1}\}},$$

which proves Theorem 2.1.

VI. Optimal Tests and Sufficient Statistics

Based on the proof of Theorem 4.1, one can design an asymptotically optimal distributed detection system as follows. Each sensor collects n observations and computes the order-k_n type (empirical distribution) corresponding to these observations. If this type is within distance η of the null distribution P (restricted to $\mathcal{X}^k$ or $\mathcal{Y}^k$ as appropriate), the sensor accepts H_0; it rejects it otherwise. Thus, each sensor communicates to the central detector a single binary message (acceptance or rejection of H_0) and the central detector accepts H_0 if and only if both received messages indicate acceptance.

In proving Theorem 4.1 we treated k_n as constant in n, which resulted in a detection scheme with error exponent equal to D_k. Yet the optimal error exponent equals $\sup_k D_k$, which may not be achieved for a finite value of k. In such cases it is necessary to take $k_n \to \infty$ at a suitable rate (details omitted) and the complexity of local encoding increases dramatically. Thus compared to the optimal conventional (non-distributed) first-order Markov detection scheme which employs the second-order ($k = 2$) empirical distribution as sufficient statistic, the decentralized scheme is considerably more complex.

As we mentioned briefly in Section II, it is tempting to attribute this difference in complexity to the fact that the marginal (X and Y) processes are not, in general, first-order Markov. Assuming that this is the case, one might offer the following explanation for the added complexity of the decentralized scheme: since each sensor essentially performs a hypothesis test on observations that are not first-order Markov, second-order types cannot possibly be sufficient. In particular, if the data are not Markov of any order, then no finite-order type can be sufficient.

Although the above argument is intuitively appealing, it does not address situations in which the marginal processes are first-order Markov. One might speculate that in such special cases, second-order types are sufficient, or equivalently, $\sup_k D_k = D_2$. A straightforward analysis shows that this is indeed true for the simplest nontrivial such case, namely testing H_0: $P = P_X \times P_Y$ versus H_1: $Q = Q_X \times Q_Y$ where each of P_X, P_Y, Q_X, and Q_Y are first-order Makov. But in other cases, the determination of

$$D_k = \min_{\tilde{P}\in\mathcal{L}_k} D(\tilde{W}_k\,\|\,V_k)$$

is not as straightforward. This problem is equivalent to finding the Markov I-projection [7] of the conditional distribution $V_k(z_k \mid z_1^{k-1}) = V(z_k \mid z_{k-1})$ on the linear space $\mathcal{L}_k$. The constraints defining $\mathcal{L}_k$ are (cf. (4.2))

$$\begin{aligned} &(\forall x_1^k) \quad \sum_{y_1^k} \tilde{P}(x_1^k, y_1^k) = P(x_1^k) \quad \text{and} \\ &(\forall y_1^k) \quad \sum_{x_1^k} \tilde{P}(x_1^k, y_1^k) = P(y_1^k). \end{aligned} \tag{6.1}$$

Using (6.1) above in conjunction with (27)–(29) in [7], one can deduce that the sought I-projection $\mu^{(k)}$ has the general form

$$\begin{aligned} \mu^{(k)}(x_1^k, y_1^k) = r(x_k, y_k)s(x_k, y_k) \\ \cdot V(x_k, y_k \mid x_1^{k-1}, y_1^{k-1})f(x_1^k)g(y_1^k) \end{aligned}$$

where the functions r, s, f, and g are not, in general, decomposable into simpler blocks. The product of the first three terms on the right-hand side is consistent with first-order Markovity, but the last two terms are not unless they have additional structure. We are thus led to believe that in the general case, the solution $\mu^{(k)}$ is not first-order Markov.

To support the above, we constructed a parametric class of examples in which the X and Y processes are both individually and jointly first-order Markov, and computed the value of D_k for $k = 2$ and $k = 3$ (the problem becomes prohibitively complex for $k > 3$). Details of the construction and numerical results are given in Appendix A. As a general rule, we found that D_3 is greater than D_2 by an appreciable margin, confirming that μ_3 is not first-order Markov. We also conjecture that in the same class of examples, D_k is strictly increasing to $\sup_k D_k$ and thus the error exponent is not achieved by any finite-order Markov scheme.

In conclusion, asymptotically optimal distributed tests on first-order Markov sources employ (in general) empirical distributions of order higher than two. This is true even in situations where the observations of individual sensors are first-order Markov, and where *ipso facto* each sensor would only need an empirical distribution of order two if it were to perform a locally (in a spatial sense) optimal test. Thus for detection of Markov sources, a distributed system can be strictly more complex than a non-distributed one. We should

note that this was not true of the i.i.d. sources treated in [3] and [4], for which asymptotically optimal tests—both centralized and distributed—could be constructed using first-order types only.

VII. Concluding Remarks

The sequence of tests described in the previous section (with $k_n \to \infty$) is universally asymptotically optimal for every level $\epsilon \in (0, 1)$ and alternative Markov distribution Q with strictly positive transition matrix V (the value of the error exponent, of course, depends on V). If the irreducibility assumption on W is relaxed to allow non-ergodic Markov sources under the null hypothesis, then the optimal tests wil have to be modified to take into account all irreducible classes. These modifications will in general depend on ϵ and on the values of $M_{X,n}$ and $M_{Y,n}$.

With minor changes, our results extend to hypothesis testing for higher-order Markov sources, yielding similar conclusions. Namely, optimal tests for order-$(k-1)$ Markov sources involve (in general) empirical distributions of order higher than k, and this is also true in cases where the marginal observations are themselves order-$(k-1)$ Markov. We have not considered stationary ergodic sources that are not finite-order Markov; progress in this direction is linked with better understanding of nondistributed hypothesis testing for general ergodic processes. A more immediate issue is what happens when the alternative transition matrix V is not strictly positive. The answer to this question is not known even in the simpler i.i.d. case.

VIII. Appendix A

We present examples in which both P and its restrictions on $\mathcal{X}^{\mathbf{Z}}$ and $\mathcal{Y}^{\mathbf{Z}}$ are first-order Markov, yet D_3 is strictly greater than D_2. This implies that the error exponent $\sup_k D_k$ is not achieved by a first-order Markov joining of P_X and P_Y.

1) The Model: For simplicity we consider binary observations, i.e., $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. The stationary Markov measures $P = \prod_i W_i$ and $Q = \prod_i V_i$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ are to be specified through P_2 and Q_2, their respective restrictions to $\mathcal{Z}^2$. The associated transition matrices $W = W_2$, $V = V_2$ and stationary distributions $W_1 = P_1$, $V_1 = Q_1$ are then easily derived. In what follows, P and Q will be drawn from the same model, so duplication of equations will not be necessary.

We restrict our model to distributions P with the local Markov property

$$P(x_2 \mid x_1 y_1) = P(x_2 \mid x_1), \quad P(y_2 \mid x_1 y_1) = P(y_2 \mid y_1). \tag{A.1}$$

The above conditions imply that the X and Y components of P are also first-order Markov. Indeed,

$$\begin{aligned} P(x_n \mid X_1^{n-1}) &= E[P(x_n \mid X_1^{n-1} Y_1^{n-1}) \mid X_1^{n-1}] \\ &= E[P(x_n \mid X_{n-1} Y_{n-1}) \mid X_1^{n-1}] \\ &= E[P(x_n \mid X_{n-1}) \mid X_1^{n-1}] \\ &= P(x_n \mid X_{n-1}), \end{aligned}$$

and similarly $P(y_n \mid Y_1^{n-1}) = P(y_n \mid Y_{n-1})$.

To construct P_2, we start with $|\mathcal{Z}|^2 - 1 = 15$ parameters, which we reduce to 12 using the stationarity constraints

$$\sum_{z_1} P_2(z_1, z) = \sum_{z_2} P_2(z, z_2). \tag{A.2}$$

Condition (A.1) imposes four additional constraints that reduce the number of free parameters to eight. Of these parameters, four may be used to specify the X and Y components of P:

$$r_x = P(X_1 = 0,\ X_2 = 0),$$

$$s_x = P(X_1 = 0,\ X_2 = 1) = P(X_1 = 1,\ X_2 = 0),$$

$$r_y = P(Y_1 = 0,\ Y_2 = 0),$$

$$s_y = P(Y_1 = 0,\ Y_2 = 1) = P(Y_1 = 1,\ Y_2 = 0).$$

Further simplification of the model is possible by assuming that the X and Y processes are interchangeable, i.e.,

$$\begin{aligned} &P(X_1 = a_1,\ Y_1 = b_1,\ X_2 = a_2,\ Y_2 = b_2) \\ &\quad = P(X_1 = b_1,\ Y_1 = a_1,\ X_2 = b_2,\ Y_2 = a_2). \end{aligned} \tag{A.3}$$

This yields

$$r_x = r_y = r, \quad s_x = s_y = s$$

and imposes two additional constraints. The free parameters are thus reduced to four, which without loss of generality can be taken as r, s (defined previously) and

$$\begin{aligned} t &= P(X_1 = 0,\ Y_1 = 0); \\ \alpha &= P(X_1 = 0,\ Y_1 = 0,\ X_2 = 0,\ Y_2 = 1). \end{aligned}$$

After some algebra, we obtain for P_2 the matrix at the bottom of the page where the entries $p_i(r, s, t, \alpha)$ are obtainable from the row sums

$$\begin{aligned} &P(X_1 = 0,\ Y_1 = 0) = t \\ &P(X_1 = 0,\ Y_1 = 1) = P(X_1 = 1,\ Y_1 = 0) = r + s - t \\ &P(X_1 = 1,\ Y_1 = 1) = 1 - 2r - 2s + t. \end{aligned}$$

$x_1y_1 \backslash x_2y_2$	00	01	10	11
00	$\frac{r}{r+s}t - \alpha$	α	α	$p_1(r, s, t, \alpha)$
01	α	$\frac{r(r+s-t)}{r+s} - \alpha$	$\frac{s(r+s-t)}{1-r-s} - \alpha$	$p_2(r, s, t, \alpha)$
10	α	$\frac{s(r+s-t)}{1-r-s} - \alpha$	$\frac{r(r+s-t)}{r+s} - \alpha$	$p_2(r, s, t, \alpha)$
11	$p_1(r, s, t, \alpha)$	$p_2(r, s, t, \alpha)$	$p_2(r, s, t, \alpha)$	$p_3(r, s, t, \alpha)$

We further observe that the distribution P_2 is time-reversible, i.e.,

$$P(X_1 = a_1,\ Y_1 = b_1,\ X_2 = a_2,\ Y_2 = b_2) = P(X_1 = a_2,\ Y_1 = b_2,\ X_2 = b_1,\ Y_2 = a_1).$$

This property can also be derived independently using (A.1)–(A.3) and the fact that $\mathcal{X}$ and $\mathcal{Y}$ are both binary. It extends to the entire measure P:

$$P(z_1, z_2, \cdots, z_{n-1}, z_n) = P(z_n, z_{n-1}, \cdots, z_2, z_1). \quad \text{(A.4)}$$

2) Evaluation of D_2 and D_3: By definition,

$$D_k = \min_{\tilde{P} \in \mathcal{L}_k} D(\tilde{W}_k \,\|\, V_k)$$

where $\mathcal{L}_k$ is the convex family of stationary measures on $\mathcal{Z}^k$ with X and Y components that coincide with those of P. We observe that $D(\tilde{W}_k \,\|\, V_k)$ is a strictly convex function of $\tilde{P}$. Indeed, it can be written as

$$\sum_{z_1^k} \tilde{P}(z_1^k) \log \frac{1}{V(z_k \mid z_{k-1})} + \sum_{z_1^k} \tilde{P}(z_1^k) \log \frac{\tilde{P}(z_1^k)}{\tilde{P}(z_1^{k-1})/4} - \log 4$$

where the first sum is a linear function of $\tilde{P}$ and the second is an unconditional divergence which strictly convex in $\tilde{P}$.

For a given $\tilde{P}$ in $\mathcal{L}_k$, let $\tilde{P}'$ and $\tilde{P}''$ represent the measures obtained by interchanging X with Y and by time-reversal, respectively. If P and Q belong to the model developed in (a) above, then

i) both $\tilde{P}'$ and $\tilde{P}''$ lie in $\mathcal{L}^k$;

ii) $D(\tilde{W}_k \,\|\, V_k) = D(\tilde{W}_k' \,\|\, V_k) = D(\tilde{W}_k'' \,\|\, V_k)$.

Using a standard convexity argument, we conclude that if $\tilde{P}$ coincides with the unique distribution $\mu^{(k)}$ that achieves D_k, then

$$\mu^{(k)} = \tilde{P} = \tilde{P}' = \tilde{P}''.$$

The above property greatly simplifies the search for $\mu^{(k)}$. For $k = 2$, the number of free parameters in $\mu^{(k)}$ is only four, while for $k = 3$, we need 16 parameters. This number is likely to grow exponentially with k, but the exact dependence is not known to us.

3) Results: We have, using the notation and results established in the proof of Theorem 4.1,

$$D_2 = D(\omega_2^{(2)} \| V_2) \le D(\omega_2^{(3)} \| V_2) \le D(\omega_3^{(3)} \| V_3) = D_3 \quad \text{(A.5)}$$

where both inequalities are equalities if and only if $\mu^{(3)}$ is first-order Markov.

We also consider D_*, defined as

$$D_* = \min_{\tilde{P} \in \mathcal{L}_*} D(\tilde{W}_2 \,\|\, V_2)$$

where $\mathcal{L}_*$ is the nonconvex class of stationary first-order Markov distributions $\tilde{P}$ on $\mathcal{Z}^{\mathbf{Z}}$ which satisfy

$$\tilde{P}(X_1, X_2, X_3) = P(X_1, X_2, X_3) \quad \text{and} \quad \tilde{P}(Y_1, Y_2, Y_3) = P(Y_1, Y_2, Y_3). \quad \text{(A.6)}$$

Clearly $\mathcal{L}_* \subset \mathcal{L}_3$, and thus

$$D_3 \le D_*. \quad \text{(A.7)}$$

Here again equality is achieved if and only if $\mu^{(3)}$ is first-order Markov.

Using the optimization package CONSOLE [13], we computed D_2, D_3, and D_* for many pairs (P, Q) from the model developed earlier and found that the inequalities in (A.5) and (A.7) are in most cases strict. Hence, $\mu^{(3)}$ is not, in general, first-order Markov.

For a specific example, we consider P specified by $r = 2/5$, $s = 1/5$, $t = 2/5$ and $\alpha = 3/80$. The corresponding P_2 matrix is given (in 240ths) by

$x_1y_1 \backslash x_2y_2$	00	01	10	11
00	55	9	9	23
01	9	23	15	1
10	9	15	23	1
11	23	1	1	23

A simple choice for Q is $Q(XY) = P(\overline{XY})$ where overbar denotes binary complement. The resulting Q_2 matrix is the reflection of P_2 about the antidiagonal and is obtained for $r = 1/5$, $s = 2/5$, $t = 1/5$, and $\alpha = 1/240$.

We found (in nats)

$$D_2 = 8.773 \times 10^{-2}$$

$$D(\omega_2^{(3)} \| V_2) = 8.775 \times 10^{-2}$$

$$D_3 = 8.830 \times 10^{-2}$$

$$D_* > 1.131 \times 10^{-1}.$$

Thus, the inequalities of (A.5) and (A.7) are decidedly strict. Consistent with the large gap between D_2 and D_* was our the observation that the distribution μ_2 achieving D_2 failed to satisfy constraint (A.6) by a margin of up to 6%.

IX. Appendix B

Proof of Lemma 3.2: We let $\mathcal{Z} = \{1, \cdots, M\}$ and $P = \prod_i W_i$ where $W_2 = W$ is irreducible and aperiodic with stationary distribution $W_1 = \pi$. On a suitable probability space, we construct an i.i.d. vector-valued process U_1^∞ with alphabet $\mathcal{U} = \mathcal{Z}^{M+1}$ as follows. We take $M + 1$ mutually independent and individually i.i.d. $\mathcal{Z}$-valued scalar processes

$$S_1^\infty, (T(1))_1^\infty, \cdots, (T(M))_1^\infty$$

with marginal distributions given by

$$(\forall r, m \in \mathcal{Z}) \quad \Pr\{S_i = m\} = \pi(m), \quad \Pr\{T_i(r) = m\} = W(m \mid r) \quad \text{(B.1)}$$

and we let

$$U_i = (S_i, T_i(1), \cdots, T_i(M)).$$

We denote the distribution of U_1^∞ by P_U.

Next we define a process Z_1^∞ by the recursion

$$Z_1 = S_1,$$

$$(n \geq 2) \qquad Z_n = T_n(Z_{n-1}).$$

From (B.1) and the above definition, it readily follows that

$$\Pr\{Z_1^n = z_1^n\} = \pi(z_1)\prod_{i=2}^{n} W(z_i \mid z_{i-1}) = P(z_1^n)$$

so that Z_1^∞ is stationary first-order Markov with distribution $P_Z = P$. The recursive definition of Z_1^∞ also implies the existence of functions $\{f_n,\ n \geq 1\}$ and $\{g_n,\ n \geq 1\}$ such that

$$Z_1^n = f_n(U_1^n) \quad \text{and}$$

$$(1 \leq i < n) \qquad Z_n = g_{n-i}(Z_i, U_{i+1}^n).$$

The i.i.d. process U_1^∞ has the blowing-up property [2]. This means that given a set A_n which contains the random sequence U_1^n with subexponentially decaying probability, it only takes "a few" (i.e., k_n, with k_n/n asymptotically vanishing) changes in U_1^n in order for A_n to contain the modified sequence with probability that approaches unity. To establish the same property for the Markov process Z_1^∞, we will show that, with sufficiently high probability, a few changes in the i.i.d. sequence U_1^n will only induce a few changes in the Markov sequence

$$Z_1^n = f_n(U_1^n).$$

1) The Positive Case: We first consider the case in which the transition matrix W has at least one column with positive entries, i.e.,

$$(\exists m)(\forall r) \qquad W(m \mid r) > 0.$$

In this case, if we let

$$\hat{\mathcal{U}} = \{(s, t(1), \cdots, t(M)) \in \mathcal{U}:\ t(1) = \cdots = t(M)\}$$

and $\overline{\mathcal{U}} = \mathcal{U} \backslash \hat{\mathcal{U}}$, then

$$p \stackrel{\text{def}}{=} P_U(\overline{\mathcal{U}}) < 1. \tag{B.2}$$

To illustrate the argument, consider a sequence u_1^n, where

$$u_i = (s_i, t_i(1), \cdots, t_i(M))$$

and let $\tilde{u}_1^n$ differ from u_1^n in the ith position only. Let the corresponding z-sequences be $z_1^n = f_n(u_1^n)$ and $\tilde{z}_1^n = f_n(\tilde{u}_1^n)$. From the definition of f_n, it is clear that the first $i-1$ positions of z_1^n will be unaffected by the change in the ith position of u_1^n (i.e., $z_1^{i-1} = \tilde{z}_1^{i-1}$) whereas, in general, $z_i^n \neq \tilde{z}_i^n$. However, the resulting error in $\tilde{z}_1^n$ can only propagate as far as the first position $j > i$ such that

$$t_j(1) = \cdots = t_j(M),$$

or equivalently, $u_j \in \hat{\mathcal{U}}$. This is so because

$$z_j = t_j(z_{j-1}) = t_j(\tilde{z}_{j-1}) = \tilde{z}_j$$

and for $j' > j$,

$$z_{j'} = g_{j'-j}(z_j, u_{j+1}^{j'}) = g_{j'-j}(\tilde{z}_j, \tilde{u}_{j+1}^{j'}) = \tilde{z}_{j'}.$$

An analogous conclusion can be drawn in the case in which $\tilde{u}_1^n$ differs from u_1^n in more than one position: the error in $\tilde{z}_1^n$ due to each change can only persist as far as the next position where all t-components are equal.

Based on the above observation we will estimate the probability that U_1^n lies in

$$G_{n,k,l} = \{u_1^n:\ (\exists \tilde{u}_1^n \in \Gamma^k\{u_1^n\})\ f_n(\tilde{u}_1^n) \notin \Gamma^{k+l-1}\{f_n(u_1^n)\}\},$$

i.e., there exist k or fewer positions in U_1^n that can induce a change in $k+l$ or more positions in $f_n(U_1^n)$.

We first define a *null run* as a finite sequence from $\overline{\mathcal{U}}$. We then let

$$G'_{n,k,l} = \{u_1^n:\ u_1^n \text{ contains, among others, } k \text{ disjoint null runs of total length } l\}.$$

From the discussion of the two previous paragraphs it follows that $G_{n,k,l}$ is a subset of $G'_{n,k,l}$.

To estimate the probability of $G'_{n,k,l}$, we first write it as the union of all sets $G''_{n,a,b}$ defined by

$$G''_{n,a,b} = \mathcal{U}^{a_1} \times \overline{\mathcal{U}}^{b_1} \times \mathcal{U}^{a_2} \times \overline{\mathcal{U}}^{b_2} \times \cdots \times \mathcal{U}^{a_k} \times \overline{\mathcal{U}}^{b_k} \times \mathcal{U}^{a_{k+1}}$$

where $\{a_i, b_i\}$ are nonnegative integers such that

$$a_1 + \cdots + a_{k+1} = n - l, \quad b_1 + \cdots + b_k = l.$$

To upper-bound the number N of distinct sets $G''_{n,a,b}$, we observe that there are $\binom{n-l+k}{k}$ choices for the vector $(a_1, \cdots, a_{k+1})$ and $\binom{l+k-1}{k-1}$ choices for the vector $(b_1, \cdots, b_k)$. Hence,

$$N \leq \binom{n-l+k}{k}\binom{l+k-1}{k-1}.$$

Actually, the above inequality is strict, since some duplications will occur between cases in which an a_i or a b_i equals 0.

From the definition of $G''_{n,a,b}$ it follows that $P_U(G''_{n,a,b}) = (P_U(\overline{\mathcal{U}}))^l = p^l$ and thus

$$P_U(G_{n,k,l}) \leq P_U(G'_{n,k,l}) \leq \binom{n-l+k}{k}\binom{l+k-1}{k-1}p^l. \tag{B.3}$$

Now consider the set $B_n \subset \mathcal{Z}^n$ in the hypothesis of the lemma and let $A_n = f_n^{-1}(B_n)$. From the definition of $G_{n,k,l}$, we obtain

$$f_n(\Gamma^k(A_n \cap G_{n,k,l}^c)) \subset \Gamma^{k+l}B_n$$

and thus

$$P_U(\Gamma^k(A_n \cap G^c_{n,k,l})) \le P_Z(\Gamma^{k+l}B_n). \tag{B.4}$$

We claim that the lemma will be established if integer sequences $\{k_n\}$ and $\{l_n\}$ are found such that:

i) the resulting set $G_{n,k,l}$ satisfies, for all sufficiently large n,

$$P_U(G_{n,k,l}) \le \frac{1}{2}\exp[-n\delta_n]; \tag{B.5}$$

ii) $k_n/(n\sqrt{\delta_n}) \to \infty$; (B.6)

iii) $(k_n + l_n)/n \to 0$. (B.7)

Indeed, in the presence of the hypothesis

$$P_U(A_n) = P_Z(B_n) \ge \exp[-n\delta_n],$$

condition (B.5) implies that

$$P_U(A_n \cap G^c_{n,k,l}) \ge \frac{1}{2}\exp[-n\delta_n].$$

A version of the blowing-up lemma for i.i.d. sequences [2, Corollary 5.3] in conjunction with condition (B.6) yields

$$P_U(\Gamma^k(A_n \cap G^c_{n,k,l})) \to 1.$$

By virtue of (B.4), the above implies that

$$P_Z(\Gamma^{k+l}B_n) \to 1$$

which in the presence of (B.7) becomes the sought conclusion.

In the final step, we demonstrate that $k_n = \lceil n\delta_n^{1/3} \rceil$ and $l_n = \lfloor n\delta_n^{1/6} \rfloor$ satisfy conditions (B.5)–(B.7) given above. We take without loss of generality $\delta \le 1$, so that $k \le l$; and we assume that $n\delta_n \ge \epsilon_n$ where $\epsilon_n \downarrow 0$ is suitably chosen. The last assumption is permissible, since the conclusion of the lemma follows trivially from the hypothesis if $n\delta_n \to 0$. For simplicity, we will also omit the subscript n from k_n, l_n, δ_n, and ϵ_n.

Since $\delta \to 0$, conditions (B.6) and (B.7) are satisfied. To investigate condition (B.5), we first consider the bound in (B.3). The binomial coefficients can be upper-bounded by

$$\binom{n}{k} \le \exp\left[nh\left(\frac{k}{n}\right)\right] \le \left(\frac{ne}{k}\right)^k$$

where $h(p) \stackrel{\text{def}}{=} -p\log p - (1-p)\log(1-p)$. The first inequality can be obtained from a type size bound (see, e.g., the proof of Lemma 5.1 in [2]); while the second inequality follows from $-\ln(1-p) \le p/(1-p)$. We thus have

$$\begin{aligned} P_U(G_{n,k,l}) &\le e^{2k-1}\left(\frac{n-l+k}{k}\right)^k\left(\frac{l+k-1}{k-1}\right)^{k-1}p^l \\ &\le e^{2k}\left(\frac{n}{k}\right)^k\left(1+\frac{l}{k}\right)^k p^l. \end{aligned}$$

Using the chosen values for k and l, we obtain

$$\begin{aligned} \frac{1}{n}\log P_U(G_{n,k,l}) \le (\delta^{1/3}+n^{-1})[2\log e + \log\delta^{-1/3} \\ + \log(\delta^{-1/6}+1)] + \delta^{1/6}\log p. \end{aligned}$$

Taking $\epsilon = n^{-2}$ (so that $n^{-1} \le \delta^{1/3}$) and invoking the inequality $\ln x^\alpha \le \alpha(x-1)$ for $\alpha > 0$, we obtain the simpler bound

$$\frac{1}{n}\log P_U(G_{n,k,l}) \le (4\log e)\delta^{1/3}[1+3\delta^{-1/12}] + \delta^{1/6}\log p.$$

The first (positive) summand on the right-hand side is $O(\delta^{1/4})$, the second [negative, by (B.2)] summand is proportional to $\delta^{1/6}$, so for suitable $c > 0$ and all sufficiently large n,

$$\frac{1}{n}\log P_U(G_{n,k,l}) \le -c\delta^{1/6} \le -\delta^{1/3} - \delta \le -\frac{1}{n} - \delta.$$

This establishes (B.5) and completes our proof.

2) The General Case: If no column of the transition matrix W is entirely positive, the above argument is clearly inapplicable because $p = 1$. Consider the following modification.

Since W is irreducible and aperiodic, the process Z_1^∞ is strongly mixing. This implies that the n-step transition probability from state m to state r converges to $\pi(r) > 0$ as $n \to \infty$ and thus there exists $d > 0$ such that any state can be reached from any other state in exactly d transitions. In particular, there exists a collection of M allowable paths, each originating from a different state m and terminating at state 1 after d transitions:

$$\begin{array}{lccccc} \text{Path 1:} & z_{10} & z_{11} & z_{12} & \cdots & z_{1d} \\ \text{Path 2:} & z_{20} & z_{21} & z_{22} & \cdots & z_{2d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Path } M\text{:} & z_{M0} & z_{M1} & z_{M2} & \cdots & z_{Md} \end{array}.$$

Here $z_{m0} = m$ and $z_{md} = 1$ for all m. We may (if necessary) modify the above collection by a simple recursion to ensure that whenever two paths meet, they merge. In other words,

$$z_{mj} = z_{rj} \Rightarrow (\forall j' > j) \quad z_{mj'} = z_{rj'}. \tag{B.7}$$

Relationship (B.7) allows us to embed the above paths in a sequence $\hat{u}_1^d$ of nonzero probability, such that whenever $\hat{u}_1^d$ occurs in the U-process, the derived Z-process is certain to be driven to state 1, i.e.,

$$(\forall m) \qquad g_d(m, \hat{u}_1^d) = 1. \tag{B.8}$$

To construct $\hat{u}_1^d$, we first write it out as the array

$$\begin{array}{cccc} s_1 & s_2 & \cdots & s_d \\ t_1(1) & t_2(1) & \cdots & t_d(1) \\ t_1(2) & t_2(2) & \cdots & t_d(2) \\ \vdots & \vdots & \ddots & \vdots \\ t_1(M) & t_2(M) & \cdots & t_d(M) \end{array}.$$

Then for each path $z_{m1}, \cdots, z_{md}$ in turn, we assign values to one entry per column j using the recursion

$$(1 \le j \le d) \qquad t_j(z_{m,j-1}) = z_{mj}.$$

Relationship (B.7) ensures that no inconsistencies will arise if an entry is visited in more than one recursion (i.e., by more than one path). For entries that are left unassigned, we choose values that have nonzero probability, e.g., $t_j(r) = \arg\max_m W(m \mid r)$. It is easy to check that (B.8) is satisfied and that $P_U\{\hat{u}_1^d\} > 0$. We let $p \stackrel{\text{def}}{=} 1 - P_U\{\hat{u}_1^d\}$.

To complete the proof, we redefine a null run as any finite sequence from $\mathcal{U}$ that does not contain the string $\hat{u}_1^d$, and we denote the set of null runs of length b by $\overline{\mathcal{U}}_b$. Retaining the previous definitions of $G_{n,k,l}$ and $G'_{n,k,l}$, we again have $G_{n,k,l} \subset G'_{n,k,l}$ and $G'_{n,k,l} = \bigcup_{a,b} G''_{n,a,b}$. In this case $G''_{n,a,b}$ is given by

$$G''_{n,a,b} = \mathcal{U}^{a_1} \times \overline{\mathcal{U}}_{b_1} \times \mathcal{U}^{a_2} \times \overline{\mathcal{U}}_{b_2} \times \cdots \times \mathcal{U}^{a_k} \times \overline{\mathcal{U}}_{b_k} \times \mathcal{U}^{a_{k+1}}$$

where the integers $\{a_i, b_i\}$ are constrained as before. The probability of each $G''_{n,a,b}$ can be upper bounded as follows. If u_1^b is a null run and $c = \lfloor b/d \rfloor$, then none of the consecutive substrings $u_1^d, u_{d+1}^{2d}, \cdots, u_{(c-1)d+1}^{cd}$ equals $\hat{u}_1^d$. Thus,

$$P_U(G''_{n,a,b}) = \prod_{i=1}^{k} P_U(\overline{\mathcal{U}}_{b_i}) \le \prod_{i=1}^{k} p^{\lfloor b_i/d \rfloor} \le p^{(l/d)-k}.$$

Using the same upper bound on the number N of sets $G''_{n,a,b}$ as before, we obtain

$$\begin{aligned} P_U(G_{n,k,l}) &\le P_U(G'_{n,k,l}) \\ &\le \binom{n-l+k}{k} \binom{l+k-1}{k-1} p^{(l/d)-k} \end{aligned}$$

where $p < 1$. The above bound is asymptotically as good as (B.3) if $k = \delta^{1/3}$ and $l = \delta^{1/6}$, and the proof can be completed as before. △

REFERENCES

[1] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 360–365, May 1985.

[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, "Multiterminal detection with zero-rate data compression," *IEEE Trans. Inform. Theory*, vol. 38, Mar. 1992.

[5] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1991.

[6] 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.

[7] I. Csiszár, T. M. Cover, and B-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788–801, Nov. 1987.

[8] K. Marton and P. Shields, "Ergodic processes and zero divergence," presented at the *1991 IEEE Int. Symp. Inform. Theory*, Budapest, Hungary, June 23–28, 1991.

[9] K. Marton and P. Shields, "The positive-divergence and blowing-up properties," *Israel J. Math.*, to be published.

[10] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer, 1984.

[11] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\bar{d}$ distance with applications to information theory," *Ann. Probab.*, vol. 3, pp. 315–328, 1975.

[12] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.

[13] M. K. H Fan, L.-S. Wang, J. Koninckx, and A. L. Tits, "Software package for optimization-based design with user-supplied simulators," *IEEE Contr. Syst. Mag.*, vol. 9, no. 1, Jan. 1989.