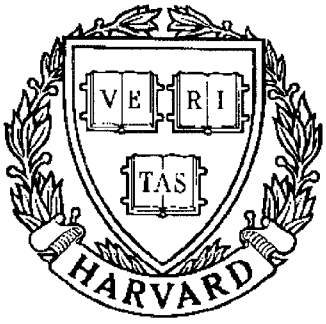
THESIS REPORT
Ph.D.

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

Distributed Hypothesis Testing with Data Compression

by H.M.H. Shalaby
Advisor: A. Papamarcou

Ph.D. 91-3

ABSTRACT

Title of Dissertation: DISTRIBUTED HYPOTHESIS TESTING WITH DATA COMPRESSION

Hossam M. H. Shalaby, Doctor of Philosophy, 1991

Dissertation directed by: Dr. Adrian Papamarcou,
Assistant Professor,
Electrical Engineering Department and
Systems Research Center

We evaluate the performance of several multiterminal detection systems, each of which comprises a central detector and a network of remote sensors. The function of the sensors is to collect data on a random signal source and process this information for transmission to the central detector. Transmission is via noiseless channels of limited capacity, hence data compression is necessary for each sensor. Upon receipt of the transmitted information, the central detector seeks to determine whether the true distribution governing the signal source belongs to a null class Π or an alternative class Ξ. System optimization is effected under the classical criterion that stipulates minimization of the type II error rate subject to an upper bound ϵ on the type I error rate. We consider the asymptotic performance—measured by an appropriate error exponents—of five types of systems. The first type has a fixed number of sensors, and processes spatially dependent but temporally independent data of growing sample size in time. Data compression for this type is at rate that tends to zero, and distribution classes Π and Ξ each consist of a single element. The second type of system is identical to the first, except for the classes Π and Ξ, which are composite.

The third type of system is a variant of the first which employs fixed-rate data compression. The fourth type is altogether different, in that it employs a variable number of sensors handling independent data of fixed sample size, and inter-sensor communication is effected by two distinct feedback schemes. The fifth type of system is yet another variant of the first in which data exhibit Markovian dependence in time and are compressed by fixed-bit quantizers. In the majority of cases we obtain concise characterizations of the associated error exponents using information-theoretic tools.

DISTRIBUTED HYPOTHESIS TESTING WITH DATA COMPRESSION

by

Hossam M. H. Shalaby

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in Partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1991

Advisory Committee:

Assistant Professor Adrian Papamarcou, Chairman/Advisor
Associate Professor Evaggelos Geraniotis
Associate Professor Prakash Narayan
Associate Professor Steven Tretter
Professor Abram Kagan

DEDICATION

To my dear parents

ACKNOWLEDGMENT

Praise to the Lord who endowed me with the circumstances to fulfill this work.

I would like to express my gratitude to my advisor, Dr. A. Papamarcou for his valuable guidance, creative suggestions, and constant encouragement throughout this work.

I appreciate the useful comments and suggestions of Dr. T. S. Han, as well as, the members of the advisory committee, Drs. E. Geraniotis, P. Narayan, S. Tretter, and A. Kagan.

I also would like to thank my parents for their encouragement to pursue my goals; and my wife for her patience, sacrifice, and moral support without which this work would not have been completed.

This research was supported by the Systems Research Center, University of Maryland, College Park, through the National Science Foundation's Engineering Research Centers Program.

TABLE OF CONTENTS

NOTATION AND CONVENTIONS

$\mathbf{N}$	set of natural numbers $\{1, 2, \ldots\}$
$\mathbf{Z}$	set of all integers $\{0, \pm 1, \ldots\}$
$\mathcal{U}$, $\mathcal{V}$, $\mathcal{W}$, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$	denote finite alphabets
U, V, W, X, Y, Z	random variables ranging over above alphabets
X^n	random vector $(X_1, \ldots, X_n)$
$X \to Y \to Z$	X, Y, Z form a Markov chain in this order
$\mathcal{X}^n$	cartesian product of n copies of $\mathcal{X}$
A^c	complement of set A
$\|A\|$	number of elements (cardinality) of a set A
x^n	n-dimensional vector $(x_1, \ldots, x_n)$
x_i^j	vector $(x_i, \ldots, x_j)$, where $j \geq i$
$\mathcal{P}(\mathcal{X})$	space of all probability distributions on $\mathcal{X}$
P_X, $\bar{P}_X$, $\tilde{P}_X$, $\hat{P}_X$, $\check{P}_X$	probability distributions on $\mathcal{X}$
Q_X, $\tilde{Q}_X$	
1_X	set of degenerate distributions in $\mathcal{P}(\mathcal{X})$
$\|\| \cdot \|\|$	denotes sup norm
$P_X \ll \bar{P}_X$	P_X is absolutely continuous w.r.t. $\bar{P}_X$
$\mathcal{B}_\eta(P_X)$	ball of distributions in $\mathcal{P}(\mathcal{X})$ centered at P_X and having sup norm radius η
P_X^n	product of n copies of P_X: $P_X^n(x^n) = \prod_{i=1}^n P_X(x_i)$
$P_X \times P_Y$	product measure on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X on $\mathcal{X}$ and P_Y on $\mathcal{Y}$
V_k, W_k	stochastic matrices on $\mathcal{Z} \times \mathcal{Z}^{k-1}$

π_W	stationary distribution of W_k (if exists and is unique)
$P = \pi_W \diamond W_k$	stationary $(k-1)$th order Markov measure on the Borel field of $\mathcal{Z}^{\mathbf{Z}}$ with stationary distribution π_W and transition matrix W_k (p. 95)
$\lfloor r \rfloor$	largest integer not exceeding r
$\lceil r \rceil$	smallest integer not less than r
$a \vee b$	larger of numbers a and b
$a \wedge b$	smaller of numbers a and b
$d_H(\cdot,\cdot)$	Hamming distance (p. 19)
$\Gamma^k(A)$	Hamming k neighborhood (p. 19)
exp, log	understood to base 2
$a \log \frac{a}{b}$	equals 0 if $a = 0$
$H(X)$	entropy functional
$H(Y\|X)$	conditional entropy
$I(X \wedge Y)$	mutual information
$I(X \wedge Y\|Z)$	conditional mutual information
$D(P\|\|Q)$	informational divergence
$D(V\|\|W\|P)$	conditional divergence
$d(P_X, P_Y\|\|Q_{XY})$	$\min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY}\|\|Q_{XY})$
$\lambda_{z^n} = \lambda_z$	type of a sequence z^n (p. 15)
$\mathcal{P}_n(\mathcal{Z})$	set of all types of sequences in $\mathcal{Z}^n$ (p. 15)
$\hat{T}_Z^n = \hat{T}_Z$	set of sequences z^n of type $\hat{P}_Z$ (p. 15)
$\tilde{T}_{Z,\eta}^n = \tilde{T}_{Z,\eta}$	set of $(\tilde{P}_Z, \eta)$-typical sequences in $\mathcal{Z}^n$ (p. 16)
$\lambda_{z_1^n}^{(k)} = \lambda_z^{(k)}$	kth order type of a sequence z^n (p. 98)

$\mathcal{P}_n^{(k)}(\mathcal{Z})$	set of all kth order types of sequences in $\mathcal{Z}^n$ (p. 98)
$\hat{T}_Z^{(k)}$	set of sequences z^n of kth order type $\hat{P}_k$ (p. 98)
$\stackrel{\text{def}}{=}$	equal by definition
$\triangle$	end of proof

CHAPTER 1

INTRODUCTION

In a signal detection problem one typically encounters two or more signal categories or states of nature, and a set of random observations. Solving the problem entails finding an optimum way of processing the observations so as to decide on the true state of nature. The problem can be formulated as an M-ary hypothesis testing problem: based on n data samples, one seeks a decision rule for declaring which of the M competing hypotheses is true. There are many useful ways of defining optimality, most notably the classical, Bayes, and minimax formulations.

The signal detection problems considered in this thesis are binary hypothesis testing problems ($M = 2$) with the classical optimality criterion. We thus have two hypotheses H_0 (the null hypothesis) and H_1 (the alternative hypothesis), and we assume that the distribution of the n data samples belongs to one of two disjoint classes of distributions Π (under H_0) and Ξ (under H_1). Acceptance of H_1 when H_0 is true is termed *type I error* or *false alarm*, while acceptance of H_0 when H_1 is true is termed *type II error* or *miss*. In conformity with the classical criterion, we seek a test (i.e., decision rule) which minimizes the probability of type II error subject to a prescribed upper bound on the probability of type I error. This upper bound is called the *significance level*, or simply *level*, of the test. The optimum test in this case is well known and is given by the Neyman-Pearson lemma [1].

In *simple* binary hypothesis testing, the classes Π and Ξ consist of single

distributions P and Q respectively. For the case in which the n data samples are independent and identically distributed according to marginal distributions P_X and Q_X, the exponential rate of decrease (in n) of the least type II error probability achievable under the classical criterion is given by Stein's lemma [8]. This rate of convergence, called the *error exponent*, is shown by that lemma to be equal to $D(P_X||Q_X)$ for all nontrivial choices of significance level. Here $D(\cdot||\cdot)$ is the Kullback-Leibler *informational divergence* [2,3] defined for distributions on a finite space $\mathcal{X}$ by

$$D(P_X||Q_X) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)} \ .$$

Stein's lemma illustrates a connection between statistics and information theory in that it characterizes the asymptotics of classical hypothesis testing by means of a functional that is firmly grounded in the information-theoretic literature. As it turns out, it is also possible to give proofs of that lemma using arguments that are almost exclusively information-theoretic. This interface between the two disciplines has also been explored by Kullback and Leibler [2,3], who extensively used the divergence functional to study hypothesis testing problems; and by Csiszár [6], who applied some statistical techniques developed by Sanov [4] and Hoeffding [5] to information theory.

Traditionally, in formulating signal detection problems one usually assumes that the data are collected by sensors and communicated fully to the decision maker or central detector. This assumption is not, however, always realistic. One often encounters situations in which either the communication channel between sensors and decision maker is of restricted capacity, or else system design dictates

that the cost and burden of information processing is shared by sensors and central detector alike. In this case the original observations are not communicated fully to the central detector; instead, they are compressed by local encoders prior to transmission to the central detector. This compression clearly results in loss of information; no further loss is assumed, i.e., transmission to the central detector is assumed error-free. Thus the central detector in effect observes the outputs of the local encoders, and processes these into a binary decision as to the true hypothesis. We note that in the case of a system comprising a central detector and one sensor, it is possible for the sensor to transmit a single binary digit without loss of optimality. This is because the encoder can perform the task of the central detector in declaring which hypothesis is true, and thereafter transmit the (binary) decision. The situation clearly changes if the system comprises more than one sensor; such a system is commonly known as a *distributed*, or *decentralized* detection system.

Tenney and Sandell were the first to address problems in distributed detection. In their pioneering work [18], they considered a setup consisting of a *fusion center* (i.e., decision maker) linked to two remote sensors. They assumed independent observations across sensors, with each sensor transmitting a binary local decision to the fusion center. Upon receipt of the local decisions, the fusion center decides which hypothesis is true using a prescribed decision rule. The contribution of that work was the optimal design, on the basis of a Bayesian criterion, of the local decision rules for an *a priori* fixed fusion rule. It was shown that the local decision rule for each sensor is given by a likelihood ratio test with an appropriate threshold.

In the same setup, Chair and Varshney [20] studied the optimum fusion rule for prescribed local decision rules. They showed that the optimum fusion rule can be evaluated using a weighted sum of the binary decisions of the sensors. The weights are functions of the probability of miss and probability of false alarm of the individual sensors.

A similar problem with N independent sensors was studied in [21], under a classical criterion for the local and global (i.e., fusion) decisions. The optimum fusion rule, for *a priori* fixed local decision rules, was obtained. Moreover, a variant of the above problem was studied in which each sensor transmits, in addition to its binary decision, quality information bits that indicate its confidence on that decision. It was shown by examples that these quality bits can lead to considerable improvement in the performance of the overall system.

In [28], Geraniotis and Chau considered a multi-sensor detection system with observations characterized by incomplete knowledge of the underlying probability distributions. The minimax robust fusion rules were derived for three different fusion schemes, namely block, sequential, and serial fusion rules. The robust decision rules were shown to utilize threshold tests based on likelihood ratios at both the sensors and the fusion center. Moreover, under the assumption of identical sensor statistics and large number of sensors, it was shown that the optimality of the sequential fusion rule remains unaffected if the thresholds employed by the local decision rules are constrained to be equal.

A plethora of interesting issues in distributed detection have been discussed in the literature; selected references include [14,15,19,22–30,32,33]. It should be noted that most known results assume that the observations are conditionally

independent (given the hypotheses) across the sensors, which is not always true in practice. Yet once the independence assumption is removed, decentralized detection problems become quite difficult from an algorithmic viewpoint. This was pointed out in a study by Tsitsiklis and Athans [22] of the complexity of computation of optimal schemes for such problems, where it was also suggested that in most practical applications one should seek heuristic or *ad hoc* designs that achieve "good" results, and forgo the pursuit of absolute optimality. Instances of studies involving dependence in time and/or across sensors include [27], where a known weak signal in m-dependent or ϕ-mixing noise is considered. Each sensor employs (suboptimal) decision tests based on memoryless nonlinearities. Based on minimizing an average cost function, the optimal nonlinearities are determined for the case of two sensors.

The models encountered in the aforementioned studies employed local encoders that compressed their observations into messages having a fixed range of values independent of sample size n. The information-theoretic approach to the distributed detection problem pursued in [9–13] broadened the class of admissible encoders by allowing codebooks (i.e., sets of messages) whose size varied with n. It also yielded complete results on the asymptotic properties of optimal distributed detection systems with spatially dependent observations. It is this approach that we intend to follow in this thesis.

In its simplest form, our setup comprises a detector or decision maker linked to two remote sensors S_X and S_Y. The sensors S_X and S_Y observe the respective components of the random sequence $\{(X_i, Y_i)\}_{i=1}^{n}$, and encode their observations using a maximum of $nR_X(n)$ and $nR_Y(n)$ bits, respectively. Thus $R_X(n)$ and

$R_Y(n)$ are the *rates* of the local encoders. Upon receipt of the two codewords, the detector accepts or rejects H_0 in accordance with the classical criterion that stipulates minimization of the type II error probability subject to a fixed upper bound ϵ on the type I error probability.

It is worth noting that one particular model, namely that for which the observed sequence $\{(X_i, Y_i)\}_{i=1}^n$ is temporally independent but spatially dependent and $R_X(n) = R_Y(n) = \infty$, admits a rather simple analysis. In that case, neither sensor needs to compress its data, and the detector knows the observed sequence $\{(X_i, Y_i)\}_{i=1}^n$ precisely. The optimal decision rule is then specified by the Neyman-Pearson lemma; and using Stein's lemma, the resulting error exponent is given by

$$D(P_{XY}||Q_{XY}) ,$$

where P_{XY} (resp. Q_{XY}) denotes the distribution of (X_i, Y_i) under H_0 (resp. H_1).

In the case of arbitrary $R_X(n)$ and $R_Y(n)$, the determination of the optimal detector is a highly complex task that also involves the optimization of the data encoders at S_X and S_Y. For this reason, it is preferable to study tractable compression/decision schemes which are *asymptotically* optimal, i.e., achieve the same error exponents as their optimal counterparts. The investigations in [10-13] are examples of such studies.

In [10], Ahlswede and Csiszár discussed the problem of one-sided fixed-rate compression (i.e., $R_X(n) = R_X$ and $R_Y(n) = \infty$). In the special case of hypothesis testing against independence (i.e., $Q_{XY} = P_X \times P_Y$), they obtained a single-letter characterization (i.e., computable characterization) of the error exponent by recourse to entropy characterization techniques [17]. Also, in the

general case where $Q_{XY} > 0$, they showed that the error exponent is independent of the value of the upper bound ϵ on the probability of type I error. Yet the problem of single-letter characterization of the error exponent in the case $Q_{XY} \neq P_X \times P_Y$ remained unsolved; single-letter lower bounds to that exponent were obtained in both [10] and [11] using compression/decision schemes whose asymptotic optimality was not established.

Amari and Han [13] considered this problem from a differential geometric viewpoint. They restricted their attention to classes of encoders which were based on symmetric functions and emulated discrete memoryless channels. Their results showed that the bounds given by Han [11] were indeed achievable by use of the above types of encoders. In a somewhat different model involving exponentially decaying bounds on the type I error rate, Han and Kobayashi [12] developed good lower bounds on the error exponent for one- and two-sided fixed-rate compression.

Han [11] studied also the case of one-bit compression, i.e., $nR_X(n) = 1$ and/or $nR_Y(n) = 1$. In this case he was able to find an asymptotically optimal compression scheme and evaluate the corresponding error exponent.

This thesis builds on the aforementioned contributions by applying information theoretic tools to the asymptotic study of distributed detection systems involving (i) variable sample sizes or numbers of sensors; (ii) spatial and/or temporal data dependence; and (iii) fixed or variable-rate data compression. The material is organized as follows.

In Chapter 2 we study the problem of simple hypothesis testing under one- and two-sided compression at asymptotically zero rate, i.e., $R_X(n) \to 0$ and/or

$R_Y(n) \to 0$. This is motivated by the one-bit compression problems studied in [11]. Observations are assumed to be i.i.d. under both hypotheses and correlated across sensors. A complete characterization of the error exponent, under a positivity constraint on the alternative distribution, is obtained. It is also shown that this error exponent is independent of the upper bound ϵ on the type I error probability and is insensitive to variations in compression rate as long as the asymptotic rate on at least one of the sensors is zero. Extensions to three or more sensors under analogous assumptions are also obtained.

In Chapter 3 we study *composite* hypothesis testing under compression at asymptotically zero rate. We obtain characterizations of the error exponents associated with four different composite hypothesis problems. These characterizations demonstrate that the conclusions of Chapter 1 do not hold in the more general situation of hypotheses involving composite distribution classes, in that the error exponent depends on both the level ϵ and the actual compression rate.

Chapter 4 is devoted to one-sided compression at fixed rate. We find a sequence of lower bounds on the error exponent and show that this sequence converges to the true value. A complete (single-letterization) solution is given for the special case of testing against a product alternative $Q_{XY} = Q_X \times Q_Y$.

In Chapter 5 we consider decentralized detection systems with a large number N of sensors that collect spatially independent and identically distributed data. Each sensor transmits a fixed number of bits to the fusion center. This system was studied in [23] where it was shown that under a Bayesian criterion the sensors can use the same decision rule without loss of asymptotic ($N \to \infty$) optimality. In this chapter we investigate the effects of two types of feedback

on the above system. One entails the transmission of sensor data to the fusion center in two stages with broadcast of feedback information from the center to the sensors after the first stage. The other involves information exchange between sensors prior to transmission to the fusion center. We show that under the classical Neyman-Pearson criterion, only the latter type of feedback yields an improvement on the asymptotic performance (in terms of the error exponent) of the above system as $N \to \infty$.

In Chapter 6 we remove the memoryless (i.i.d.) assumption on the information source and extend the results obtained in Chapter 2 to stationary ergodic Markov sources. Specifically, we obtain the error exponent for simple hypothesis testing under compression with fixed codebook sizes.

Before proceeding to the main body of our results, we would like to give a brief overview of the method of attack followed in this work. Most of our results are characterizations of error exponents. In proving the main theorems, we use the common technique of information theory whereby both a *direct* (or *positive*) and a *converse* result are established. In the direct result, we construct a sequence of encoding schemes that achieves the figure of merit (i.e., error exponent) proposed in the statement of the theorem. In the converse part, we show that we cannot find an encoding scheme that yields a better error exponent. The main tools used in establishing direct results involve the concept of a typical sequence [17], which is discussed in Chapter 2. Most converse results in this thesis depend on a pivotal theorem (also given in Chapter 2) whose proof is based on the celebrated blowing-up lemma [17].

CHAPTER 2

MULTITERMINAL SIMPLE HYPOTHESIS TESTING WITH ZERO-RATE DATA COMPRESSION

2.1. Introduction

We consider the problem of testing a simple null hypothesis H_0 against a simple alternative H_1 on the basis of compressed data from a discrete-time, discrete-alphabet, memoryless multiple source. In its simplest form, our setup comprises two remote *sensors* S_X and S_Y which are linked to a *central detector*. The sensors S_X and S_Y observe the respective components of the random sequence $\{(X_i, Y_i)\}_{i=1}^{n}$, and encode their observations into a maximum of M_n and N_n messages, respectively. Upon receipt of the two codewords, the central detector accepts or rejects the null hypothesis in conformity with the classical criterion that stipulates minimization of the probability of falsely accepting H_0 (*type II error*) subject to a fixed upper bound ϵ on the probability of falsely rejecting H_0 (*type I error*).

Distributed detection systems of the above type have been widely studied in the recent literature. The models most frequently encountered [18–30] employ fixed codebook sizes $M_n = M$ and $N_n = N$, where M and N are often equal to 2. In such cases, the central detector receives from each sensor what amounts to a local decision, possibly accompanied by an assessment (on a fixed finite-valued scale) of the sensor's confidence in that decision. Of course, it is also possible to design distributed detection systems employing varying codebook sizes M_n and N_n, as is the case with certain models discussed in the information-theoretic

literature [9–13,31] and in this thesis.

As we discussed in Chapter 1, the particular model in which M_n and N_n are large enough so that no compression is needed, the analysis is well known. In that case the optimal decision rule for testing $H_0 : P_{XY}$ versus $H_1 : Q_{XY}$ at any level ϵ is specified by the Neyman-Pearson lemma. Furthermore, the resulting minimum type II error probability $\beta_n(\epsilon)$ satisfies the asymptotic identity

$$-\lim_n \frac{1}{n} \log \beta_n(\epsilon) \;=\; D(P_{XY}||Q_{XY}) \;.$$

The quantity appearing on the left-hand side of the above equation (which is due to Stein [8]) is termed the *error exponent* for the hypothesis testing problem.

In this chapter we consider the hypothesis testing problem under data compression at (asymptotically) *zero-rate*. In other words, we assume that the codebook sizes satisfy constraints of the type

$$R_X(n) \;=\; \frac{1}{n} \log M_n \;\to\; 0, \qquad R_Y(n) \;=\; \frac{1}{n} \log N_n \;\to\; 0.$$

Our inquiry was motivated by the study in [11] of hypothesis testing under two-sided one-bit ($M_n = 2,\ N_n = 2$), and one-sided one-bit ($M_n = 2,\ N_n = \infty$), compression. For those systems, Han proposed a simple scheme that compressed both S_X and S_Y to one bit, was independent of the level ϵ, and yielded a simply characterized lower bound on the error exponent. He then proved by converse theorems the tightness of the lower bound

(i) for all values of ϵ in the case of two-sided one-bit compression;

(ii) for a range $(0, \epsilon_0)$ of values of ϵ, where $\epsilon_0 < 1$, in the case of one-sided one-bit compression.

We complement and extend the above results as follows. For fixed-level simple hypothesis testing under the positivity constraint $Q_{XY} > 0$, we prove that the two-sided one-bit compression/decision scheme proposed by Han in [11] is, for *all* $\epsilon \in (0,1)$, asymptotically optimal in the broader class of one-sided zero-rate compression/decision schemes. Thus an optimal distributed detection system employing two sensors, of which one transmits data at a vanishing rate while the other supplies complete information about its observations, is asymptotically no better than optimal system in which each sensor transmits a single binary digit. It also follows as a special case that optimal systems for fixed codebook-size compression ($M_n = M$, $N_n = N$) have the same asymptotic performance regardless of the values M and N. In other words, no gain in asymptotic performance can result by allowing each sensor to transmit a quantized, or *soft*, decision [19,21,30] instead of a binary, or *hard*, decision.

The formulation of the general problem is given in Section 2.2, together with pertinent notation. The converse theorem for simple hypothesis testing appears in Section 2.3, followed in Section 2.4 by an extension to the multivariate case (r sensors, where $r > 2$). Section 2.5 contains some concluding remarks.

2.2. Problem statement and preliminaries

(a) General notation. The observations of S_X and S_Y are denoted by the sequences $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ and $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$, respectively, and the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are assumed finite. Since the multiple source is memoryless, the sequence of pairs $((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ is i.i.d. under both hypotheses. In what follows, it will be convenient to deal with the product space $\mathcal{X}^n \times \mathcal{Y}^n$ instead of $(\mathcal{X} \times \mathcal{Y})^n$, and thus the observations will be

collectively represented by the pair $(X^n, Y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.

By virtue of the aforementioned i.i.d. assumption, all distributions of interest can be specified through bivariate distributions on $\mathcal{X} \times \mathcal{Y}$. Under the null hypothesis, the distribution of any pair (X_i, Y_i) is usually denoted by P_{XY}, and its respective marginals by P_X and P_Y. The distributions of X^n, Y^n, and (X^n, Y^n) under the same hypothesis are denoted by P_X^n, P_Y^n and P_{XY}^n, respectively. The i.i.d. assumption then implies that for all (x^n, y^n) in $\mathcal{X}^n \times \mathcal{Y}^n$,

$$P_{XY}^n(x^n, y^n) = \prod_{i=1}^{n} P_{XY}(x_i, y_i) \ .$$

Analogous notation is employed for the alternative hypothesis, with Q replacing P. We will also have occasion to use distributions $\bar{P}_{XY}$, $\tilde{P}_{XY}$ and $\hat{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, which will yield marginals and higher-order distributions in the same manner as P_{XY} and Q_{XY}.

The spaces of all distributions on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$ will be denoted by $\mathcal{P}(\mathcal{X})$, $\mathcal{P}(\mathcal{Y})$ and $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, respectively.

The compression of X^n and Y^n is effected by encoders f_n, and g_n, respectively, where

$$f_n : \mathcal{X}^n \mapsto \{1, \ldots, M_n\} \ , \qquad \text{and} \qquad g_n : \mathcal{Y}^n \mapsto \{1, \ldots, N_n\} \ .$$

For one-sided zero-rate compression of X^n we assume that $N_n \geq |\mathcal{Y}|^n$ and

$$M_n \ \geq \ 2 \ , \qquad \lim_n \frac{1}{n} \log M_n = 0 \ , \tag{2.1}$$

and similarly for one-sided zero-rate compression of Y^n, we have $M_n \geq |\mathcal{X}|^n$ and

$$N_n \ \geq \ 2 \ , \qquad \lim_n \frac{1}{n} \log N_n = 0 \ . \tag{2.2}$$

For two-sided zero-rate compression, both (2.1) and (2.2) are assumed.

The central detector is represented by the function

$$\phi_n : \{1, \ldots, M_n\} \times \{1, \ldots, N_n\} \mapsto \{0, 1\} \ ,$$

where the output 0 signifies the acceptance of the null hypothesis H_0, and 1 its rejection. This induces a partition of the original (i.e., non-compressed) sample space $\mathcal{X}^n \times \mathcal{Y}^n$ into an *acceptance* region

$$\mathcal{A}_n \stackrel{\text{def}}{=} \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \phi_n(f_n(x^n), g_n(y^n)) = 0\} \ ,$$

and a *critical* (or rejection) region $\mathcal{A}_n^c$.

By nature of the encoding process, the acceptance region can be decomposed into M_n rectangles $C_i \times F_i$ in $\mathcal{X}^n \times \mathcal{Y}^n$ that possess disjoint projections C_i on $\mathcal{X}^n$. More precisely, if for every $1 \leq i \leq M_n$ we define

$$C_i \ = \ \{x^n \in \mathcal{X}^n : f_n(x^n) = i\} \qquad \text{and} \qquad F_i \ = \ \{y^n \in \mathcal{Y}^n : \phi_n(i, g(y^n)) = 0\} \ ,$$

then we can write

$$\mathcal{A}_n \ = \ \bigcup_{i=1}^{M_n} C_i \times F_i \ , \qquad \text{where} \qquad (\forall i \neq j) \quad C_i \cap C_j = \emptyset \ . \tag{2.3}$$

We can obtain an alternative representation for $\mathcal{A}_n$ by partitioning $\mathcal{Y}^n$ into N_n sets:

$$\mathcal{A}_n \ = \ \bigcup_{i=1}^{N_n} D_i \times G_i \ , \qquad \text{where} \qquad (\forall i \neq j) \quad D_i \cap D_j = \emptyset \ . \tag{2.4}$$

Note that conditions (2.3) and (2.4) *jointly* characterize all admissible acceptance regions under *two-sided* compression with codebook sizes M_n (for X^n) and N_n (for Y^n). Taken *separately*, the above conditions characterize the admissible acceptance regions under *one-sided* compression of X^n and Y^n, respectively.

(b) Simple hypothesis testing. The optimal acceptance region for testing $H_0 : P$ versus $H_1 : Q$ at a given level $\epsilon \in (0,1)$ is one that minimizes $Q_{XY}^n(\mathcal{A}_n)$ over all acceptance regions $\mathcal{A}_n$ that

(C1) yield a value of $P_{XY}^n(\mathcal{A}_n^c)$ less than or equal to ϵ;

and

(C2) satisfy the appropriate compression constraints; namely

- (2.1) and (2.3) for one-sided compression of X^n;
- (2.2) and (2.4) for one-sided compression of Y^n;
- (2.1), (2.2), (2.3) and (2.4) for two-sided compression.

The resulting *minimum* probability of type II error is denoted by $\beta_n(M_n, N_n, \epsilon)$, and the associated error exponent is given by

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \stackrel{\text{def}}{=} -\lim_n \frac{1}{n} \log \beta_n(M_n, N_n, \epsilon) \ ,$$

provided the limit on the right-hand side exists.

(c) Typical sequences. Our proofs rely on the concept of a typical sequence, as developed in [17]. We cite some basic definitions and facts on typical sequences.

The *type* of a sequence $x^n \in \mathcal{X}^n$ is the distribution λ_x on $\mathcal{X}$ defined by the relationship

$$(\forall a \in \mathcal{X}) \qquad \lambda_x(a) \stackrel{\text{def}}{=} \frac{1}{n} N(a|x^n),$$

where $N(a|x^n)$ is the number of terms in x^n equal to a. The set of all types of sequences in $\mathcal{X}^n$, namely $\{\lambda_x : x^n \in \mathcal{X}^n\}$, will be denoted by $\mathcal{P}_n(\mathcal{X})$.

Given a type $\hat{P}_X \in \mathcal{P}_n(\mathcal{X})$, we will denote by $\hat{T}_X^n$ the set of sequences $x^n \in \mathcal{X}^n$ of type $\hat{P}_X$:

$$\hat{T}_X^n \stackrel{\text{def}}{=} \{x^n \in \mathcal{X}^n : \lambda_x = \hat{P}_X\} \ .$$

Also, for an arbitrary distribution $\tilde{P}_X$ on $\mathcal{X}$ and a constant $\eta > 0$, we will denote by $\tilde{T}^n_{X,\eta}$ the set of $(\tilde{P}_X, \eta)$-*typical* sequences in $\mathcal{X}^n$. A sequence x^n is $(\tilde{P}_X, \eta)$-typical if $|\lambda_x(a) - \tilde{P}_X(a)| \leq \eta$ for every letter $a \in \mathcal{X}$ and, in addition, $\lambda_x(a) = 0$ for every a such that $\tilde{P}_X(a) = 0$. Thus, if $||\cdot||$ denotes the sup norm and $\ll$ denotes absolute continuity, we have

$$\tilde{T}^n_{X,\eta} \stackrel{\text{def}}{=} \{x^n \in \mathcal{X}^n\colon\ ||\lambda_x - \tilde{P}_X|| \leq \eta,\ \lambda_x \ll \tilde{P}_X\}\ .$$

In the same manner, we will denote by $T^n_{X,\eta}$ and $\bar{T}^n_{X,\eta}$ the sets of (P_X, η)- and $(\bar{P}_X, \eta)$- (respectively) typical sequences in $\mathcal{X}^n$. We will have no need to consider sequences with exact or approximate type Q_X.

The proofs of the following lemmas appear in [17]. As usual, $|\mathcal{A}|$ denotes the size of $\mathcal{A}$.

LEMMA 2.1. *The size of $\mathcal{P}_n(\mathcal{X})$ is at most $(n+1)^{|\mathcal{X}|}$. For any $\hat{P}_X$ in $\mathcal{P}_n(\mathcal{X})$ and Q_X in $\mathcal{P}(\mathcal{X})$,*

$$(n+1)^{-|\mathcal{X}|} \exp[nH(\hat{P}_X)]\ \leq\ |\hat{T}^n_X|\ \leq\ \exp[nH(\hat{P}_X)]\ ,$$

and

$$(n+1)^{-|\mathcal{X}|}\ \exp[-nD(\hat{P}_X||Q_X)]\ \leq\ Q^n_X(\hat{T}^n_X)\ \leq\ \exp[-nD(\hat{P}_X||Q_X)]\ .$$

LEMMA 2.2. *For any distribution P_X on $\mathcal{X}$ and $\eta > 0$,*

$$P^n_X(T^n_{X,\eta})\ \geq\ 1\ -\ \frac{|\mathcal{X}|}{4n\eta^2}\ .$$

One can easily modify the above exposition to accommodate pairs $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ by reverting to their representation in $(\mathcal{X} \times \mathcal{Y})^n$. Thus the type of (x^n, y^n) is the distribution λ_{xy} on $\mathcal{X} \times \mathcal{Y}$ such that

$$\lambda_{xy}(a,b) \;=\; \frac{1}{n}\Big|\{i \;:\; (x_i, y_i) = (a,b)\}\Big| \;,$$

and the class $\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})$, as well as the sets $\hat{T}^n_{XY} \subset \mathcal{X}^n \times \mathcal{Y}^n$ and $\tilde{T}^n_{XY,\eta} \subset \mathcal{X}^n \times \mathcal{Y}^n$, are defined accordingly.

In this and the following sections, we will omit the superscript n from T^n, as n will be essentially constant.

2.3. A converse theorem for simple hypothesis testing.

In this section, we derive the error exponent for simple hypothesis testing under the positivity condition $Q_{XY} > 0$. We show that the error exponent $\theta(\mathbf{M}, \mathbf{N}, \epsilon)$ is independent of ϵ and the compression scheme used (one-sided or two-sided), provided the asymptotic zero-rate constraints (2.1) and/or (2.2) are met. Furthermore, its value is given by the minimum of the quantity

$$D(\tilde{P}_{XY} \| Q_{XY})$$

over all bivariate distributions $\tilde{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ whose marginals on $\mathcal{X}$ and $\mathcal{Y}$ agree with those of P_{XY}.

The positive result, namely the existence of a sequence of acceptance regions that achieve the above value, was shown in [11]. The acceptance regions used in that work had the simple rectangular form

$$T_{X,\eta} \times T_{Y,\eta} \;,$$

and were thus admissible under the most stringent of two-sided compression schemes, namely $M_n = N_n = 2$. Our result here is a strong converse for one-sided compression of X^n, i.e., we show that for every value of $\epsilon \in (0,1)$ and every sequence of acceptance regions $\mathcal{A}_n$ satisfying (C1), (2.1) and (2.3), the following is true:

$$-\liminf_n \frac{1}{n} \log Q_{XY}^n(\mathcal{A}_n) \;\leq \min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY} || Q_{XY}) \; .$$

By symmetry, the same is true for one-sided compression of Y^n, and *a fortiori*, for two-sided compression.

THEOREM 2.1. *Let P_{XY} be arbitrary, and $Q_{XY} > 0$. For all $\epsilon \in (0,1)$ and sequences M_n and $\mathcal{A}_n$ satisfying conditions (2.1) and (2.3), the following is true: if for every n,*

$$P_{XY}^n(\mathcal{A}_n^c) \leq \epsilon \; ,$$

then

$$-\liminf_n \frac{1}{n} \log Q_{XY}^n(\mathcal{A}_n) \leq \min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY} || Q_{XY}) \; .$$

PROOF. By (2.3), we have

$$\mathcal{A}_n \;=\; \bigcup_{i=1}^{M_n} C_i \times F_i \; ,$$

where the C_i's are pairwise disjoint. Assume that $P_{XY}^n(\mathcal{A}^c) \leq \epsilon$, or equivalently, $P_{XY}^n(\mathcal{A}) \geq 1 - \epsilon$. Then there exists an index i_0 such that

$$P_{XY}^n(C_{i_0} \times F_{i_0}) \;\geq\; \frac{1-\epsilon}{M_n} \; .$$

Letting $C = C_{i_0}$ and $F = F_{i_0}$, we can rewrite the above as

$$P_{XY}^n(C \times F) \;\geq\; \exp(-n\delta_n), \tag{3.1}$$

where

$$\delta_n \;=\; \delta_n(M_n, \epsilon) \;=\; -\frac{1}{n}\log(1-\epsilon) + \frac{1}{n}\log M_n \;,$$

and $\delta_n \to 0$ by condition (2.1). Equation (3.1) clearly implies that

$$P_X^n(C) \;\geq\; \exp(-n\delta_n) \qquad \text{and} \qquad P_Y^n(F) \;\geq\; \exp(-n\delta_n) \;. \tag{3.2}$$

Thus asymptotically, neither C nor F has "exponentially small" probability. By the blowing-up lemma [17, Theorem 5.4], this fact implies that both sets possess Hamming k_n-neighborhoods which are asymptotically "as thin" as the sets themselves (i.e., $k_n/n \to 0$), and whose probabilities approach unity as n tends to infinity. Specifically, let $d_H(\cdot,\cdot)$ denote Hamming distance; that is for any $u^n,\ x^n \in \mathcal{X}^n$,

$$d_H(u^n, x^n) \;\stackrel{\text{def}}{=}\; \sum_{i=1}^{n} d_H(u_i, x_i) \;,$$

where

$$d_H(u_i, x_i) \;\stackrel{\text{def}}{=}\; \begin{cases} 0, & \text{if } u_i = x_i \;; \\ 1, & \text{otherwise} \;; \end{cases}$$

and define the Hamming k-neighborhood $\Gamma^k C$ of C by

$$\Gamma^k C \;\stackrel{\text{def}}{=}\; \{u^n \in \mathcal{X}^n \;:\; (\exists\, x^n \in C)\; d_H(x^n, u^n) \leq k\} \;.$$

The blowing-up lemma asserts that under condition (3.2), there exist sequences k_n and γ_n satisfying

$$k_n/n \to 0 \qquad \text{and} \qquad \gamma_n \to 0 \;,$$

and such that

$$P_X^n(\Gamma^{k_n} C) \;\geq\; 1-\gamma_n \qquad \text{and} \qquad P_Y^n(\Gamma^{k_n} F) \;\geq\; 1-\gamma_n \;. \tag{3.3}$$

Furthermore, k_n and γ_n depend only on $|\mathcal{X}|$, $|\mathcal{Y}|$ and δ_n, and *not* on P_{XY}.

In what follows, we will use k instead of k_n in all superscripts.

Equation (3.3) clearly holds true if we replace P by $\tilde{P}$, where $\tilde{P}_{XY}$ satisfies the marginal constraints

$$\tilde{P}_X = P_X \qquad \text{and} \qquad \tilde{P}_Y = P_Y \ .$$

Using the elementary property $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$, we then obtain

$$\tilde{P}^n_{XY}(\Gamma^k C \times \Gamma^k F) \ \geq \ \tilde{P}^n_X(\Gamma^k C) \ + \ \tilde{P}^n_Y(\Gamma^k F) \ - \ 1 \ ,$$

and hence

$$\tilde{P}^n_{XY}(\Gamma^k C \times \Gamma^k F) \ \geq \ 1 - 2\gamma_n \ . \tag{3.4}$$

Thus under the n-fold product of $\tilde{P}_{XY}$, the probability of the rectangle $\Gamma^k C \times \Gamma^k F$ approaches unity as n tends to infinity. By Lemma 2.2, the same is true of the set of $(\tilde{P}_{XY}, \eta)$-typical elements in $\mathcal{X}^n \times \mathcal{Y}^n$, where $\eta = \eta_n = n^{-1/3}$. Indeed,

$$\tilde{P}^n_{XY}(\tilde{T}_{XY,\eta}) \ \geq \ 1 - \frac{|\mathcal{X}||\mathcal{Y}|}{4n\eta_n^2} \ = \ 1 - \frac{|\mathcal{X}||\mathcal{Y}|}{4n^{1/3}} \ .$$

Hence, for all sufficiently large n, we obtain

$$\tilde{P}^n_{XY}((\Gamma^k C \times \Gamma^k F) \cap \tilde{T}_{XY,\eta}) \ \geq \ \frac{1}{2} \ . \tag{3.5}$$

By definition of $\tilde{T}_{XY,\eta}$, we have the following decomposition:

$$\tilde{T}_{XY,\eta} = \bigcup_{\substack{\hat{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}): \\ ||\hat{P}_{XY} - \tilde{P}_{XY}|| \leq \eta, \\ \hat{P}_{XY} \ll \tilde{P}_{XY}}} \hat{T}_{XY} \ .$$

Thus, observing that the elements of a given $\hat{T}_{XY}$ are equiprobable under any i.i.d. measure, we can rewrite (3.5) as

$$\sum_{\substack{\hat{P}_{XY}\in\mathcal{P}_n(\mathcal{X}\times\mathcal{Y}):\\ ||\hat{P}_{XY}-\tilde{P}_{XY}||\le\eta,\\ \hat{P}_{XY}\ll\tilde{P}_{XY}}} \tilde{P}_{XY}^n(\hat{T}_{XY})\,\frac{|(\Gamma^k C\times\Gamma^k F)\cap\hat{T}_{XY}|}{|\hat{T}_{XY}|} \;\ge\; \frac{1}{2}\,.$$

At least one of the fractions in the above sum must be greater than or equal to $1/2$; hence there exists a type $\hat{P}_{XY}\in\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})$ satisfying

$$||\hat{P}_{XY}-\tilde{P}_{XY}||\;\le\;\eta \qquad\text{and}\qquad \hat{P}_{XY}\ll\tilde{P}_{XY}\;,$$

and such that

$$\frac{|(\Gamma^k C\times\Gamma^k F)\cap\hat{T}_{XY}|}{|\hat{T}_{XY}|}\;\ge\;\frac{1}{2}\,.$$

Since pairs (x^n,y^n) of the same type are also equiprobable under Q_{XY}^n, we conclude that for the above type $\hat{P}_{XY}$,

$$\begin{aligned} Q_{XY}^n(\Gamma^k C\times\Gamma^k F) &\ge Q_{XY}^n((\Gamma^k C\times\Gamma^k F)\cap\hat{T}_{XY}) \\ &= Q_{XY}^n(\hat{T}_{XY})\frac{|(\Gamma^k C\times\Gamma^k F)\cap\hat{T}_{XY}|}{|\hat{T}_{XY}|} \\ &\ge \frac{1}{2}Q_{XY}^n(\hat{T}_{XY})\,. \end{aligned} \tag{3.6}$$

We have thus established that the probabilities of the sets $\Gamma^k C\times\Gamma^k F$ and $\hat{T}_{XY}$ are of the same exponential order under Q_{XY}^n. We now show that the same is true of the pair $\Gamma^k C\times\Gamma^k F$ and $C\times F$. The argument is similar to that given in [10], Section IV.

Consider an arbitrary element (u^n,v^n) of $\Gamma^k C\times\Gamma^k F$. By definition of Γ^k, there exists at least one element $(x^n,y^n)\in C\times F$ such that (x_i,y_i) differs from (u_i,v_i) for at most $2k_n$ values of i. We thus have

$$\begin{aligned} Q_{XY}^n(u^n,v^n) &= \prod_{i=1}^n Q_{XY}(u_i,v_i) \\ &\le \rho^{-2k}\prod_{i=1}^n Q_{XY}(x_i,y_i) \;=\; \rho^{-2k}Q_{XY}^n(x^n,y^n)\,, \end{aligned} \tag{3.7}$$

where

$$\rho \stackrel{\text{def}}{=} \min_{x \in \mathcal{X},\ y \in \mathcal{Y}} Q_{XY}(x,y) > 0 \ .$$

As (u^n, v^n) ranges over $\Gamma^k C \times \Gamma^k F$, each element (x^n, y^n) of $C \times F$ will be selected at most $|\Gamma^k(x^n)| \cdot |\Gamma^k(y^n)|$ times. By virtue of this, (3.7) yields

$$Q_{XY}^n(\Gamma^k C \times \Gamma^k F) \ \leq \ \rho^{-2k} |\Gamma^k(x^n)||\Gamma^k(y^n)| Q_{XY}^n(C \times F) \ .$$

From [17] we have the upper bound

$$|\Gamma^k(x^n)| \ \leq \ \exp\Big[n\Big(h\Big(\frac{k_n}{n}\Big) + \frac{k_n}{n} \log |\mathcal{X}|\Big)\Big] \ ,$$

where $h(\cdot)$ denotes the binary entropy function. Thus we may write

$$Q_{XY}^n(\Gamma^k C \times \Gamma^k F) \ \leq \ \exp(n\xi_n) Q_{XY}^n(C \times F) \ , \tag{3.8}$$

where

$$\xi_n \ = \ 2h\Big(\frac{k_n}{n}\Big) + \frac{k_n}{n} \log(|\mathcal{X}||\mathcal{Y}|) - \frac{2k_n}{n} \log \rho \ \to \ 0 \ .$$

As a final step, we combine equations (3.6) and (3.8) with the upper bound on $Q_{XY}^n(\hat{T}_{XY})$ provided by Lemma (2.1). Thus

$$\begin{aligned} Q_{XY}^n(C \times F) &\geq \frac{1}{2} \exp(-n\xi_n) Q_{XY}^n(\hat{T}_{XY}) \\ &\geq \frac{(n+1)^{-|\mathcal{X}||\mathcal{Y}|}}{2} \exp[-n(D(\hat{P}_{XY}||Q_{XY}) + \xi_n)] \\ &\geq \exp[-n(D(\hat{P}_{XY}||Q_{XY}) + \zeta_n)], \end{aligned}$$

where

$$\zeta_n \ = \ \zeta_n(\rho, \epsilon, M_n, |\mathcal{X}|, |\mathcal{Y}|) \ \to \ 0 \ .$$

Over the range of pairs $(\tilde{P}_{XY}, \tilde{Q}_{XY})$ such that and $\tilde{Q}_{XY} \geq \rho$, the divergence functional $D(\tilde{P}_{XY}||\tilde{Q}_{XY})$ is convex and bounded, and thus also uniformly continuous. It follows that we can find a sequence

$$\mu_n \ = \ \mu_n(\rho, |\mathcal{X}|, |\mathcal{Y}|) \ \to \ 0$$

such that

$$||\hat{P}_{XY} - \tilde{P}_{XY}|| \leq \eta_n = n^{-1/3} \quad \Rightarrow \quad \left| D(\hat{P}_{XY}||Q_{XY}) - D(\tilde{P}_{XY}||Q_{XY}) \right| \leq \mu_n \ .$$

Hence

$$Q_{XY}^n(C \times F) \ \geq \ \exp[-n(D(\tilde{P}_{XY}||Q_{XY}) + \zeta_n + \mu_n)] \ , \tag{3.9}$$

and consequently

$$-\liminf_n \frac{1}{n} \log Q_{XY}^n(\mathcal{A}_n) \ \leq \ D(\tilde{P}_{XY}||Q_{XY}) \ .$$

Since $\tilde{P}_{XY}$ satisfies the appropriate marginal constraints, the proof is complete.

$\triangle$

The above result, in conjunction with the positive part of Theorem 5 in [11], yields

THEOREM 2.2. *If $Q_{XY} > 0$, the error exponent for $H_0 : P_{XY}$ versus $H_1 : Q_{XY}$ under one-sided or two-sided zero-rate compression is given by*

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \ = \min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X, \ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY}||Q_{XY}) \ . \qquad \triangle$$

REMARK. In the proof of the converse theorem, the constants ζ_n and μ_n appearing on the right-hand side of equation (3.9) are independent of the distributions P_{XY}, $\tilde{P}_{XY}$, and depend on Q_{XY} only through the lower bound ρ. With this in mind, we state without proof the following variant of Theorem 2.1, which will be useful in establishing converse results in the chapters that follow.

THEOREM 2.3. *Fix $\rho > 0$ and $\epsilon \in (0,1)$, and let M_n be a sequence of integers satisfying (2.1). Then there exists a sequence*

$$\nu_n \ = \ \nu_n(\rho, \epsilon, M_n, |\mathcal{X}|, |\mathcal{Y}|) \ \to \ 0$$

such that for every $\tilde{Q}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ *that satisfies* $\tilde{Q}_{XY} \geq \rho$, *and every* $\tilde{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $C \in \mathcal{X}^n$, $F \in \mathcal{Y}^n$ *that satisfy either*

$$(\exists P_{XY} : \ P_X = \tilde{P}_X, \ P_Y = \tilde{P}_Y) \qquad P_{XY}^n(C \times F) \ \geq \ \frac{1-\epsilon}{M_n}$$

or, more generally,

$$\tilde{P}_X^n(C) \ \geq \ \frac{1-\epsilon}{M_n} , \qquad \tilde{P}_Y^n(F) \ \geq \ \frac{1-\epsilon}{M_n} ,$$

the following is true:

$$\tilde{Q}_{XY}^n(C \times F) \ \geq \ \exp[-n(D(\tilde{P}_{XY}||\tilde{Q}_{XY}) + \nu_n] \ . \qquad \triangle$$

2.4. Arbitrary number of sensors.

The results of the previous section can be extended to multiterminal detection systems employing r sensors, where $r > 2$. Here the problem is that of testing $H_0 : P$ versus $H_1 : Q$, where P and Q are r-variate distributions. As in the case $r = 2$, we assume $Q > 0$ and that at least $r - 1$ source components are compressed at asymptotically zero rate. It is then possible to prove an analog of Theorem 2.2 stating that the error exponent is given by the minimum of $D(\tilde{P}||Q)$ over all r-variate distributions $\tilde{P}$ whose univariate marginals agree with those of P. We give a sketch of the proof for the case of three sensors S_X, S_Y, and S_Z, at least the first two of which are compressed at asymptotically zero rate.

Direct part. Following the proof of Theorem 5 [11], we propose a sequence of acceptance regions $\mathcal{A}_n$ in $\mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$ defined by

$$\mathcal{A}_n \ = \ T_{X,\eta}^n \times T_{Y,\eta}^n \times T_{Z,\eta}^n \ ,$$

where $\eta = \eta_n = n^{-1/3}$. One can easily show that $\mathcal{A}_n \supset T^n_{XYZ,\zeta}$ for suitable $\zeta = \zeta_n > 0$, $\zeta_n \to 0$. Thus by Lemma 2.2.,

$$P^n_{XYZ}(\mathcal{A}_n) \geq 1 - \epsilon$$

for all sufficiently large n. Also, by the type-counting argument given in [11], one can establish the relationship

$$-\lim_n \frac{1}{n} \log Q^n_{XYZ}(\mathcal{A}_n) \;=\; \min_{\substack{\tilde{P}_{XYZ}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y, \\ \tilde{P}_Z = P_Z}} D(\tilde{P}_{XYZ} || Q_{XYZ}) \;.$$

Converse part. Assuming that X^n and Y^n are compressed to a maximum of M_n and N_n bits, respectively, we can write any acceptance region $\mathcal{A}_n$ in $\mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$ as

$$\mathcal{A}_n \;=\; \bigcup_{i=1}^{M_n} \bigcup_{j=1}^{N_n} C_i \times E_j \times F_{ij} \;.$$

Here the F_{ij}'s are subsets of $\mathcal{Z}^n$, and the C_i's and E_j's form partitions of $\mathcal{X}^n$ and $\mathcal{Y}^n$, respectively. Thus there exists a subset $C \times E \times F$ of $\mathcal{A}_n$ such that

$$P^n_{XYZ}(C \times E \times F) \;\geq\; \frac{1-\epsilon}{M_n N_n} \;.$$

Using the asymptotic zero rate conditions (2.1) and (2.2), one obtains the counterpart of (3.1), namely

$$P^n_{XYZ}(C \times E \times F) \;\geq\; \exp(-n\delta_n) \;,$$

and the proof proceeds as before. $\triangle$

As in the bivariate case, the value of the exponent does not depend on the level ϵ and the codebook sizes. Thus in particular, systems employing one-bit compression per source component can attain the same asymptotic performance

as more complex systems employing zero-rate compression on $r-1$ source components, and no compression at all on the remaining component.

More generally, if r_c out of r source components are compressed at asymptotically zero rate and the remaining $r_u = r - r_c$ are not compressed, the error exponent can be shown (by suitably modifying the arguments in the proofs of Theorem 5, [11] and Theorem 2.1 above) to be given by the minimum of $D(\tilde{P}||Q)$ over all distributions $\tilde{P}$ that agree with P on:

(i) the univariate marginals corresponding to the compressed source components; and

(ii) the r_u-variate marginal corresponding to the r_u components which are not compressed.

Thus the latter r_u components are essentially treated as one. It also follows for $r_u \geq 2$ that if we impose zero-rate compression on any one of these r_u components, then the error exponent will (in general) decrease.

2.5. Concluding remarks

The positivity assumption on the alternative hypothesis was essential for the derivation of the converse results in this chapter. Without this assumption, we could not have applied the blowing-up lemma in the proof of the pivotal Theorems 2.1 and 2.3. The same difficulty was encountered in the proof of the converse result in [10], Theorem 6, which also employed the blowing-up lemma. We hope that this obstacle will eventually be removed.

In the meantime, we should note that in the case of simple hypothesis testing, there are instances where $Q \not> 0$ and $D(\tilde{P}||Q)$ is trivially minimized by $\tilde{P} = P$. In such cases, the resulting minimum is equal to the error exponent under no data

compression (cf. Stein's lemma [8]), and the converse result follows immediately.

We must also emphasize that Theorem 2.2 does not subsume its counterpart in [11]. Although the converse theorem appearing in that work was valid for one-bit compression of S_X and for $\epsilon \in (0, \epsilon_0)$ only, the hypothesis of that theorem did not impose any constraints on Q_{XY} other than $D(P_{XY}||Q_{XY}) < \infty$.

The technique used in the proof of Theorem 2.1 also yielded Theorem 2.3, which is the basis for the converse results on systems involving composite hypothesis testing and decision feedback, to be treated in subsequent chapters.

Finally, by considering the proof of Theorem 2.1 we observe that the i.i.d. assumption on the joint process $\{(X_i, Y_i)\}$ under the null hypothesis can be relaxed. That is, our results still hold if the above process is such that:

(i) $\{X_i\}$ and $\{Y_i\}$ are *individually* i.i.d. under P_X and P_Y, respectively; and

(ii) $\{(X_i, Y_i)\}$ is *jointly* i.i.d. under the alternative hypothesis Q_{XY}.

CHAPTER 3

MULTITERMINAL COMPOSITE HYPOTHESIS TESTING WITH ZERO-RATE DATA COMPRESSION

3.1. Introduction

In this chapter we consider issues of optimal zero-rate compression for *composite* hypothesis testing. In other words, for disjoint classes Π and Ξ of bivariate distributions on $\mathcal{X} \times \mathcal{Y}$, we wish to test

$$H_0: P_{XY} \in \Pi \qquad \text{against} \qquad H_1: Q_{XY} \in \Xi$$

subject to the compression rate constraints:

$$\lim_n \frac{1}{n} \log M_n = 0, \qquad \text{and/or} \qquad \lim_n \frac{1}{n} \log N_n = 0.$$

In the previous chapter, we studied the corresponding simple hypothesis testing problem ($|\Pi| = |\Xi| = 1$). Let us briefly recapitulate the conclusions of that chapter. Under a positivity assumption on the alternative distribution, we showed that the error exponent $\theta(\mathbf{M}, \mathbf{N}, \epsilon)$ of the minimum type II error probability exists and is independent of the sequences $\mathbf{M}$, $\mathbf{N}$ and the level ϵ. Furthermore, it is possible to specify a sequence of asymptotically optimal acceptance regions *solely* in terms of the null distribution P, and thus the alternative distribution enters the picture only in the computation of the error exponent $\theta(\mathbf{M}, \mathbf{N}, \epsilon)$.

In this chapter we ascertain that the above conclusions are of limited validity in the case of composite hypothesis testing. That is, the error exponent for the above composite hypothesis test depends in general on the sequences $\mathbf{M}$, $\mathbf{N}$, and

the level ϵ. Furthermore, the choice of optimal acceptance regions is influenced by *both* Π and Ξ.

More specifically, assuming a uniform positivity constraint on the distributions in Ξ, we show the following.

(a) If Π and Ξ are arbitrary classes, and the codebook sizes M_n, N_n are allowed to grow without bound subject to the above zero-rate constraints, then the error exponent has no further dependence on $\mathbf{M}$, $\mathbf{N}$, and ϵ, and is achieved by a sequence of acceptance regions specified solely in terms of Π.

Assuming that the null class Π is finite, and that the codebook sizes M_n and N_n are *fixed* at M and N, respectively, we also have:

(b) If no two distributions in Π share the same X or Y marginal, then the optimal acceptance regions and the resulting error exponents depend on Π, Ξ, M and N. There exist threshold values of M and N, above which we can specify optimal acceptance regions in terms of the null class Π alone.

(c) If two or more distributions in Π share the same X or Y marginal, then the solution of the problem depends explicitly on the level ϵ (in addition to Π, Ξ, M and N).

In (a) above, we consider the problem in its full generality and derive a compact expression for the error exponent. To illustrate (b), we produce a complete solution for the setup in which

$$|\Pi| = 2\,, \qquad |\Xi| = 1\,, \qquad M = 2\,,$$

and the S_Y encoder is nontrivial, i.e., $N \geq 2$. To illustrate (c), we consider the

situation in which

$$|\Pi| < \infty\,, \qquad |\Xi| = 1\,, \qquad M = 2\,,$$

and N is greater than a certain threshold. The results in (b) and (c) admit extensions to larger codebooks and classes of distributions, albeit at some expense of compactness in the characterization of the error exponent. It seems to us that the general problem of determining error exponents for arbitrary Π, Ξ, $\mathbf{M}$ and $\mathbf{N}$ resists coherent treatment, and is thus placed outside the scope of this thesis.

The formulation of the general problem is given in Section 3.2, together with pertinent notation. The main results (a), (b), and (c) appear in Sections 3.3, 3.4, and 3.5, respectively.

3.2. Problem statement and notation

(a) Composite hypothesis testing. Let Π and Ξ be disjoint subsets of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. For testing $H_0 : P_{XY} \in \Pi$ versus $H_1 : Q_{XY} \in \Xi$ at a given level ϵ, we employ the *uniformly most powerful* (UMP) test. Thus for a given level $\epsilon \in (0,1)$, we seek to minimize the quantity

$$\sup_{Q \in \Xi} Q_{XY}^n(\mathcal{A}_n)$$

over all acceptance regions $\mathcal{A}_n$ that meet the constraints

(C1) $P_{XY}^n(\mathcal{A}_n^c) \leq \epsilon$ for *all* P_{XY} in Π;

and

(C2) satisfy the appropriate compression constraints; namely

- (2.1) and (2.3) of Chapter 2 for one-sided compression of X^n;

- (2.2) and (2.4) of Chapter 2 for one-sided compression of Y^n;
- (2.1), (2.2), (2.3) and (2.4) of Chapter 2 for two-sided compression.

We use the definition

$$\beta_n(M_n, N_n, \epsilon) \stackrel{\text{def}}{=} \min_{\mathcal{A}_n} \sup_{Q \in \Xi} Q_{XY}^n(\mathcal{A}_n) \ ,$$

and define the associated error exponent as

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \stackrel{\text{def}}{=} -\lim_n \frac{1}{n} \log \beta_n(M_n, N_n, \epsilon) \ ,$$

provided the limit on the right-hand side exists.

(b) Notation. The following notation will be used in this chapter.

(i) For a class of distributions Π on $\mathcal{X} \times \mathcal{Y}$, the corresponding classes of marginals are denoted by

$$\Pi_X = \{P_X \in \mathcal{P}(\mathcal{X}) \colon \exists P_{XY} \in \Pi\}, \quad \text{and} \quad \Pi_Y = \{P_Y \in \mathcal{P}(\mathcal{Y}) \colon \exists P_{XY} \in \Pi\}.$$

(ii) In the space $\mathcal{P}(\mathcal{X})$, we define a ball of radius η centered at P_X by

$$\mathcal{B}_\eta(P_X) \stackrel{\text{def}}{=} \{\tilde{P}_X \in \mathcal{P}(\mathcal{X}) \colon \|\tilde{P}_X - P_X\| \le \eta, \ \tilde{P}_X \ll P_X\},$$

and we extend the domain of definition to subsets of $\mathcal{P}(\mathcal{X})$ in the obvious way.

(iii) If P_X, P_Y, Q_{XY} are distributions on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$, respectively, we let

$$d(P_X, P_Y || Q) \stackrel{\text{def}}{=} \min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X, \ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY} || Q_{XY}).$$

More generally, if Δ, Λ and Ξ are classes of distributions on the same spaces (respectively) as above, then

$$d(\Delta, \Lambda || \Xi) \stackrel{\text{def}}{=} \inf_{\substack{Q_{XY} \in \Xi, \\ \tilde{P}_{XY}: \ \tilde{P}_X \in \Delta, \ \tilde{P}_Y \in \Lambda}} D(\tilde{P}_{XY} || Q_{XY}).$$

(iv) Finally, if Φ is a subset of $\mathcal{P}(\mathcal{X})$, we will write

$$\bigcup_{P_X \in \Phi} T_X \qquad \text{and} \qquad \bigcup_{P_X \in \Phi} T_{X,\eta}$$

for

$$\bigcup_{\hat{P}_X \in \Phi \cap \mathcal{P}_n(\mathcal{X})} \hat{T}_X \qquad \text{and} \qquad \bigcup_{\hat{P}_X \in \mathcal{B}_\eta(\Phi) \cap \mathcal{P}_n(\mathcal{X})} \hat{T}_X \ ,$$

respectively.

3.3. Unboundedly growing codebook sizes.

We consider the composite hypothesis testing problem in which the null class Π is an arbitrary subset of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and the alternative class Ξ satisfies the uniform positivity constraint

$$\rho_{\inf} \stackrel{\text{def}}{=} \inf_{Q \in \Xi} \min_{(x,y) \in \mathcal{X} \times \mathcal{Y}} Q_{XY}(x,y) \ > \ 0 \ . \tag{3.1}$$

The above condition ensures that the convex function $D(\cdot||\cdot)$ is bounded on $\mathcal{P}(\mathcal{X} \times \mathcal{Y}) \times \Xi$ and is thus uniformly continuous. For the codebook sizes, we assume

$$\lim_n \frac{1}{n} \log M_n \ = \ \lim_n \frac{1}{n} \log N_n \ = \ 0, \qquad \text{and} \qquad \lim_n M_n \ = \ \lim_n N_n \ = \ \infty \ .$$

The above size constraint allows each of the two encoders to specify the type of the observed sequence with arbitrary accuracy. Indeed, if we let

$$a_n \ = \ \lfloor M_n^{1/|\mathcal{X}|} \rfloor,$$

then by an elementary geometrical construction we can partition $\mathcal{P}(\mathcal{X})$ into at most $a_n^{|\mathcal{X}|} \leq M_n$ cells $\mathcal{C}_i^n$ of maximum dimension (measured by sup norm) not

exceeding a_n^{-1}; clearly $a_n^{-1} \to 0$ since $M_n \to \infty$. The same is true for $\mathcal{P}(\mathcal{Y})$ with b_n replacing a_n:

$$b_n \;=\; \lfloor N_n^{1/|\mathcal{Y}|} \rfloor \;.$$

We denote the $\mathcal{P}(\mathcal{Y})$-counterpart of $\mathcal{C}_i^n$ by $\mathcal{F}_j^n$, and we write

$$C_i^n \;=\; \bigcup_{\hat{P}_X \in \mathcal{C}_i^n} \hat{T}_X, \qquad F_j^n \;=\; \bigcup_{\hat{P}_Y \in \mathcal{F}_j^n} \hat{T}_Y \;.$$

Based on the above partition, we devise a compression/decision scheme as follows. First, we require that each encoder transmit the cell index corresponding to the observed type, i.e.,

$$f_n(x^n) \;=\; i \qquad \text{iff} \qquad x^n \in C_i^n;$$

$$g_n(y^n) \;=\; j \qquad \text{iff} \qquad y^n \in F_j^n.$$

Next, we seek an acceptance region $\mathcal{A}_n \subset \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$\mathcal{A}_n \;\supset \bigcup_{P_{XY} \in \Pi} T_{XY,\eta}^n \tag{3.2}$$

for some *fixed* $\eta > 0$. This is because the above set has P_{XY}^n-probability that *uniformly* approaches unity for all $P_{XY} \in \Pi$ (by Lemma 2.2), and this automatically ensures that the type I error bound is met for every $\epsilon \in (0,1)$. We define $\mathcal{A}_n$ as the smallest union of rectangles $C_i^n \times F_j^n$ that contains

$$\bigcup_{P_{XY} \in \Pi} T_{X,\xi} \times T_{Y,\xi} \;,$$

where ξ is a multiple of η chosen so as to ensure that (3.2) holds.

Since ξ is fixed and the dimension of each C_i^n and F_j^n shrinks to zero as n approaches infinity, it is also true that for n sufficiently large,

$$\mathcal{A}_n \;\subset \bigcup_{P_{XY} \in \Pi} T_{X,2\xi} \times T_{Y,2\xi} \;.$$

By a standard argument based on the definition of typicality, we also have

$$T_{X,2\xi} \times T_{Y,2\xi} \subset \bigcup_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} \tilde{T}_{XY,\zeta} ,$$

where ζ is a fixed multiple of ξ and η. We conclude that

$$\mathcal{A}_n \subset \bigcup_{\substack{\tilde{P}_{XY}: \\ (\exists P_{XY} \in \Pi)\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} \tilde{T}_{XY,\zeta} .$$

A union bound on $Q^n(\mathcal{A}_n)$ for $Q \in \Xi$ can now be established using Lemma 2.1 and the fact that $D(\cdot||\cdot)$ is uniformly continuous on $\mathcal{P}(\mathcal{X} \times \mathcal{Y}) \times \Xi$:

$$\begin{aligned} Q(\mathcal{A}_n) &\leq |\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})| \exp[-n \inf_{\substack{\tilde{P}_{XY}: \\ (\exists P_{XY} \in \Pi)\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} (D(\tilde{P}_{XY}||Q_{XY}) - \mu'(\zeta))] \\ &\leq \exp[-n \inf_{P_{XY} \in \Pi} (d(P_X, P_Y||Q_{XY}) - \mu(\zeta))] , \end{aligned}$$

where $\mu(\zeta)$ goes to zero together with ζ (and hence also η). We therefore have

$$\beta_n(M_n, N_n, \epsilon) \leq \exp[-n \inf_{P_{XY} \in \Pi,\ Q_{XY} \in \Xi} (d(P_X, P_Y||Q_{XY}) - \mu(\zeta))] .$$

Since $\mu(\zeta)$ can be made arbitrarily small by choice of η, we conclude that

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \geq \inf_{P_{XY} \in \Pi,\ Q_{XY} \in \Xi} d(P_X, P_Y||Q_{XY}) .$$

To show the reverse inequality, we consider an admissible acceptance region $\mathcal{A}_n$. By (C2), for every distribution P_{XY} in Π, we can find a rectangle $C \times F \subset \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$P_{XY}^n(C \times F) \geq (1 - \epsilon)/M_n .$$

Applying Theorem 2.3 with $\rho = \rho_{\inf}$, we obtain a universal sequence $\nu_n \to 0$ with the property that for every $Q_{XY} \in \Xi$, $P_{XY} \in \Pi$ and $\tilde{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that $\tilde{P}_X = P_X$, $\tilde{P}_Y = P_Y$, the following is true:

$$Q_{XY}^n(\mathcal{A}_n) \geq \exp[-n(D(\tilde{P}_{XY}||Q_{XY}) + \nu_n)] .$$

We conclude that

$$\beta_n(M_n, N_n, \epsilon) \;\geq\; \exp[-n \inf_{\substack{\tilde{P}_{XY}:\ (\exists P_{XY}\in\Pi)\ \tilde{P}_X=P_X,\ \tilde{P}_Y=P_Y, \\ Q_{XY}\in\Xi}} (D(\tilde{P}_{XY}||Q_{XY}) + \nu_n)]$$

and hence

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \;\leq\; \inf_{P_{XY}\in\Pi,\ Q_{XY}\in\Xi} d(P_X, P_Y||Q_{XY})\ .$$

We thus have proved

THEOREM 3.1. *If $\Pi \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is arbitrary, $\Xi \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is such that*

$$\inf_{Q\in\Xi} \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} Q_{XY}(x,y) \;>\; 0\ ,$$

and the sequences $\mathbf{M}$, $\mathbf{N}$ *satisfy*

$$\lim_n \frac{1}{n}\log M_n \;=\; \lim_n \frac{1}{n}\log N_n \;=\; 0, \qquad \text{and} \qquad \lim_n M_n \;=\; \lim_n N_n \;=\; \infty\ ,$$

then

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \;=\; \inf_{P_{XY}\in\Pi,\ Q_{XY}\in\Xi} d(P_X, P_Y||Q_{XY})\ . \qquad \triangle$$

3.4. Fixed codebook sizes

In this and the following section we assume that the codebook sizes are fixed:

$$(\forall n) \qquad M_n = M, \qquad N_n = N\ .$$

Under the above constraint, it is no longer possible to encode the type of the observed sequences with arbitrary accuracy, and the conclusion of Theorem 3.1 does not hold in general. As we shall see, the optimal system design depends on the distribution classes Π and Ξ, the actual codebook sizes M and N, and (somewhat surprisingly) the value of the level ϵ.

Throughout the remainder of this chapter, we will assume for simplicity that the class Π is finite. As we pointed out earlier, some of our proofs admit cumbersome but straightforward generalizations to situations in which Π is infinite. However, since our aim is to highlight salient differences from the simple hypothesis testing problem, we choose to restrict our attention to the simplest possible setups.

Our first observation is that given Π finite and Ξ satisfying the uniform positivity constraint (3.1), there exist threshold values of M and N, above which the error exponent of Theorem 3.1 obtains. Indeed, if

$$M \;\geq\; |\Pi_X| + 1, \qquad N \;\geq\; |\Pi_Y| + 1 \;,$$

then the S_X encoder can specify which one (if any) of the distributions $P_X \in \Pi_X$ lies within distance η of the type of the observed sequence x^n; similarly for S_Y. This allows us to employ an acceptance region

$$\mathcal{A}_n \;=\; \bigcup_{P_{XY} \in \Pi} T_{X,\eta} \times T_{Y,\eta} \;.$$

As in the proof of the positive part of Theorem 3.1, we obtain

$$\theta(\mathbf{M}, \mathbf{N}, \epsilon) \;\geq\; \inf_{P_{XY} \in \Pi,\; Q_{XY} \in \Xi} d(P_X, P_Y || Q_{XY}) \;.$$

The converse part of Theorem 3.1 clearly suffices for this problem. We thus have

THEOREM 3.2. *If $\Pi \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is finite, $\Xi \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is such that*

$$\inf_{Q \in \Xi} \min_{(x,y) \in \mathcal{X} \times \mathcal{Y}} Q_{XY}(x,y) \;>\; 0 \;,$$

and

$$M \geq |\Pi_X| + 1, \qquad N \geq |\Pi_Y| + 1 ,$$

then

$$\theta(M, N, \epsilon) = \inf_{P_{XY} \in \Pi,\ Q_{XY} \in \Xi} d(P_X, P_Y || Q_{XY}) . \qquad \triangle$$

We now consider the situation in which either one or both codebook sizes M, N are smaller than the threshold values given in the hypothesis of Theorem 3.2. For simplicity, we will assume that Π consists of two distributions P_{XY}, $\bar{P}_{XY}$ with distinct X, Y marginals, and that the alternative hypothesis is simple, i.e., $\Xi = \{Q_{XY}\}$. The threshold values are then both equal to 3, and it clearly suffices to consider two cases: (i) $(M, N) = (2, 3)$ and (ii) $(M, N) = (2, 2)$.

We consider case (i) first.

THEOREM 3.3. *Let* $\Pi = \{P_{XY}, \bar{P}_{XY}\}$, *where* $P_X \neq \bar{P}_X$ *and* $P_Y \neq \bar{P}_Y$. *If* $Q_{XY} > 0$, *then for* $0 < \epsilon < 1$,

$$\theta(2, 3, \epsilon) = \theta^{(1)} \vee \theta^{(2)} ,$$

where

$$\theta^{(1)} \stackrel{\text{def}}{=} d(\Pi_X, \Pi_Y || Q)$$

and

$$\theta^{(2)} \stackrel{\text{def}}{=} d(P_X, P_Y || Q) \wedge d(\bar{P}_X, \bar{P}_Y || Q) \wedge \min_{\tilde{P}_X \in \mathcal{P}(\mathcal{X})} \{d(\tilde{P}_X, P_Y || Q) \vee d(\tilde{P}_X, \bar{P}_Y || Q)\}.$$

PROOF. *Positive part.* As before, we restrict our attention to encoders that group sequences of the same type together. Since $N = 3$, a sensible choice for the S_Y encoder is one that specifies whether the sequence y^n lies in $T_{Y,\eta}$, $\bar{T}_{Y,\eta}$ or $(T_{Y,\eta} \cup \bar{T}_{Y,\eta})^c$.

The choice of the S_X encoder is less straightforward. At first sight it would seem that since $M = 2$, the S_X encoder should specify whether or not the type of the observed sequence x^n is close to *either one* or *none* of the distributions P_X, $\bar{P}_X$, i.e.,

$$C_1 \;=\; T_{X,\eta} \cup \bar{T}_{X,\eta} \;, \qquad C_2 \;=\; (T_{X,\eta} \cup \bar{T}_{X,\eta})^c \;.$$

With this choice of encoders, the smallest acceptance region that satisfies the type I error constraint under both P_{XY} and $\bar{P}_{XY}$ is

$$\mathcal{A}_n^{(1)} \;=\; (T_{X,\eta} \cup \bar{T}_{X,\eta}) \times (T_{Y,\eta} \cup \bar{T}_{Y,\eta}) \;.$$

The Q^n-probability of the above set can be upper-bounded in the standard fashion (viz. proof of Theorem 3.1):

$$Q_{XY}^n(\mathcal{A}_n) \leq \exp[-n(\min_{\tilde{P}_X \in \{P_X, \bar{P}_X\}, \tilde{P}_Y \in \{P_Y, \bar{P}_Y\}} D(\tilde{P}_{XY}||Q_{XY}) - \mu(\eta))]$$

where $\mu(\eta) \to 0$ as $\eta \to 0$. This yields, since η is arbitrary small,

$$\theta(2, 3, \epsilon) \;\geq\; \theta^{(1)} \;=\; d(\Pi_X, \Pi_Y|Q) \;. \tag{4.1}$$

Another (somewhat less prominent) candidate for the S_X encoder is one that *separates* sequences of approximate type P_X from ones of approximate type $\bar{P}_X$. Since only two codewords are available, this separation entails grouping

some types in $\mathcal{P}(\mathcal{X})$ together with P_X, and the remaining types with $\bar{P}_X$. More formally, if

$$\Phi \subset \mathcal{P}(\mathcal{X}) - \mathcal{B}_\eta(P_X) - \mathcal{B}_\eta(\bar{P}_X) \qquad \text{and} \qquad \bar{\Phi} = \mathcal{P}(\mathcal{X}) - \mathcal{B}_\eta(P_X) - \mathcal{B}_\eta(\bar{P}_X) - \Phi \ ,$$

then this encoder partitions $\mathcal{X}^n$ into

$$C_1' = T_{X,\eta} \cup \bigcup_{\tilde{P}_X \in \Phi} \tilde{T}_X \qquad \text{and} \qquad C_2' = (C_1')^c = \bar{T}_{X,\eta} \cup \bigcup_{\tilde{P}_X \in \bar{\Phi}} \tilde{T}_X \ . \quad (4.2)$$

With this choice of S_X encoder (together with the S_Y encoder introduced in the beginning of the proof), the smallest acceptance region that satisfies the type I error constraint is

$$\mathcal{A}_n^{(2)} = (C_1' \times T_{Y,\eta}) \cup (C_2' \times \bar{T}_{Y,\eta})$$

Note that unlike $\mathcal{A}_n^{(1)}$, $\mathcal{A}_n^{(2)}$ does not contain $T_{X,\eta} \times \bar{T}_{Y,\eta}$ or $\bar{T}_{X,\eta} \times T_{Y,\eta}$. It does, however, contain pairs (x^n, y^n) whose marginal type λ_x is close to neither P_X nor $\bar{P}_X$.

To estimate $Q^n(\mathcal{A}_n^{(2)})$, we decompose each of C_1' and C_2' into two sets as in definition (4.2). We then treat $\mathcal{A}_n^{(2)}$ as a union of four disjoint sets, and upper-bound their Q^n-probabilities in the usual way:

$$\begin{aligned}
Q_{XY}^n(T_{X,\eta} \times T_{Y,\eta}) &\leq \exp[-n(d(P_X, P_Y||Q) - \mu(\eta))] \ , \\
Q_{XY}^n(\bar{T}_{X,\eta} \times \bar{T}_{Y,\eta}) &\leq \exp[-n(d(\bar{P}_X, \bar{P}_Y||Q) - \mu(\eta))] \ , \\
Q_{XY}^n\Big(\bigcup_{\tilde{P}_X \in \Phi} \tilde{T}_X \times T_{Y,\eta}\Big) &\leq \exp[-n(\inf_{\tilde{P}_X \in \Phi} d(\tilde{P}_X, P_Y||Q) - \mu(\eta))] \ , \\
Q_{XY}^n\Big(\bigcup_{\tilde{P}_X \in \bar{\Phi}} \tilde{T}_X \times \bar{T}_{Y,\eta}\Big) &\leq \exp[-n(\inf_{\tilde{P}_X \in \bar{\Phi}} d(\tilde{P}_X, \bar{P}_Y||Q) - \mu(\eta))] \ ,
\end{aligned}$$

where $\mu(\eta) \to 0$ as $\eta \to 0$.

Thus the error exponent associated with this choice of acceptance region is greater than or equal to the minimum of the four exponents appearing in the above bounds, namely the quantity

$$d(P_X, P_Y||Q) \ \wedge \ d(\bar{P}_X, \bar{P}_Y||Q) \ \wedge \ \inf_{\tilde{P}_X \in \Phi} d(\tilde{P}_X, P_Y||Q) \ \wedge \ \inf_{\tilde{P}_X \in \bar{\Phi}} d(\tilde{P}_X, \bar{P}_Y||Q) \ .$$

At this point we should note that by letting η shrink to zero, we have expanded the classes Φ and $\bar{\Phi}$ in the vicinity of P_X and $\bar{P}_X$ so that $\Phi \cup \bar{\Phi} = \mathcal{P}(\mathcal{X}) - \{P_X\} - \{\bar{P}_X\}$. This is justified by continuity of $d(\cdot, \cdot||Q)$, which further allows us to treat Φ and $\bar{\Phi}$ in the above expression as constituting a partition of $\mathcal{P}(\mathcal{X})$.

It remains to find that partition $\{\Phi, \bar{\Phi}\}$ of $\mathcal{P}(\mathcal{X})$ which maximizes

$$v(P_X) \ \wedge \ \bar{v}(\bar{P}_X) \ \wedge \ \inf_{\tilde{P}_X \in \Phi} v(\tilde{P}_X) \ \wedge \ \inf_{\tilde{P}_X \in \bar{\Phi}} \bar{v}(\tilde{P}_X) \ ,$$

where $v(\cdot) \stackrel{\text{def}}{=} d(\cdot, P_Y||Q)$ and $\bar{v}(\cdot) \stackrel{\text{def}}{=} d(\cdot, \bar{P}_Y||Q)$. This is easily accomplished by noting that

$$\begin{aligned} \inf_{\Phi} v(\tilde{P}_X) \ \wedge \ \inf_{\bar{\Phi}} \bar{v}(\tilde{P}_X) \ &\leq \ \inf_{\Phi} [v(\tilde{P}_X) \ \vee \ \bar{v}(\tilde{P}_X)] \ \wedge \ \inf_{\bar{\Phi}} [v(\tilde{P}_X) \ \vee \ \bar{v}(\tilde{P}_X)] \\ &= \ \inf_{\mathcal{P}(\mathcal{X})} [v(\tilde{P}_X) \ \vee \ \bar{v}(\tilde{P}_X)] \\ &= \ \inf_{\tilde{P}_X : v(\tilde{P}_X) \geq \bar{v}(\tilde{P}_X)} v(\tilde{P}_X) \ \wedge \ \inf_{\tilde{P}_X : v(\tilde{P}_X) < \bar{v}(\tilde{P}_X)} \bar{v}(\tilde{P}_X) \ . \end{aligned}$$

Thus an optimal partition consists of the sets

$$\Phi = \{\tilde{P}_X : v(\tilde{P}_X) \geq \bar{v}(\tilde{P}_X)\}, \quad \text{and} \quad \bar{\Phi} = \{\tilde{P}_X : v(\tilde{P}_X) < \bar{v}(\tilde{P}_X)\} \ ,$$

and the error exponent associated with the corresponding $\mathcal{A}^{(2)}$ is given by

$$\begin{aligned} \theta^{(2)} \ = \ & d(P_X, P_Y||Q) \ \wedge \ d(\bar{P}_X, \bar{P}_Y||Q) \\ & \wedge \ \min_{\tilde{P}_X \in \mathcal{P}(\mathcal{X})} \{d(\tilde{P}_X, P_Y||Q) \ \vee \ d(\tilde{P}_X, \bar{P}_Y||Q)\} \ . \end{aligned}$$

We conclude that $\theta(2,3,\epsilon) \geq \theta^{(2)}$, and in light of (4.1),

$$\theta(2,3,\epsilon) \geq \theta^{(1)} \vee \theta^{(2)} .$$

Converse part. For fixed n, consider an admissible acceptance region $\mathcal{A}_n$. By nature of the encoding, $\mathcal{A}_n$ can be written as

$$\mathcal{A}_n = (C_1 \times F_1) \cup (C_2 \times F_2) ,$$

where C_1 and C_2 form a partition of $\mathcal{X}^n$, and at most one of F_1, F_2 may be empty. From the type I error constraint

$$P_{XY}(\mathcal{A}_n) \geq 1-\epsilon \qquad \text{and} \qquad \bar{P}_{XY}(\mathcal{A}_n) \geq 1-\epsilon ,$$

it follows that two cases may arise.

Case 1. For i and j distinct, we have

$$P^n_{XY}(C_i \times F_i) \geq P^n_{XY}(C_j \times F_j) \qquad \text{and} \qquad \bar{P}^n_{XY}(C_i \times F_i) \geq \bar{P}^n_{XY}(C_j \times F_j).$$

This clearly implies that

$$\tilde{P}^n_X(C_i) \geq \frac{1-\epsilon}{2} \qquad \text{and} \qquad \tilde{P}^n_Y(F_i) \geq \frac{1-\epsilon}{2} ,$$

for any $\tilde{P}_X \in \Pi_X$, $\tilde{P}_Y \in \Pi_Y$. From Theorem 2.3, we obtain

$$-\frac{1}{n}\log Q^n_{XY}(C_i \times F_i) \leq d(\Pi_X, \Pi_Y || Q) + \nu_n = \theta^{(1)} + \nu_n ,$$

where $\nu_n \to 0$ as $n \to \infty$, and thus also

$$-\frac{1}{n}\log Q^n(\mathcal{A}_n) \leq \theta^{(1)} + \nu_n . \tag{4.3}$$

Case 2. For i and j distinct, we have

$$P_{XY}^n(C_i \times F_i) \;\geq\; P_{XY}^n(C_j \times F_j) \qquad \text{and} \qquad \bar{P}_{XY}^n(C_i \times F_i) \;<\; \bar{P}_{XY}^n(C_j \times F_j)\;. \tag{4.4}$$

Using Theorem 2.3 once again, we obtain respectively

$$-\frac{1}{n}\log Q_{XY}^n(C_i \times F_i) \;\leq\; d(P_X, P_Y||Q) + \nu_n$$

and

$$-\frac{1}{n}\log Q_{XY}^n(C_j \times F_j) \;\leq\; d(\bar{P}_X, \bar{P}_Y||Q) + \nu_n\;.$$

Hence

$$-\frac{1}{n}\log Q^n(\mathcal{A}_n) \;\leq\; d(P_X, P_Y||Q) \;\wedge\; d(\bar{P}_X, \bar{P}_Y||Q) + \nu_n\;. \tag{4.5}$$

Relationship (4.4) also implies that

$$P_Y^n(F_i) \;\geq\; \frac{1-\epsilon}{2} \qquad \text{and} \qquad \bar{P}_Y^n(F_j) \;\geq\; \frac{1-\epsilon}{2}\;.$$

By virtue of Theorem 2.3, the above inequalities can lead to a further upper bound on $Q^n(\mathcal{A}_n)$ provided there exists a distribution $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$ for which either $\tilde{P}_X^n(C_i)$ or $\tilde{P}_X^n(C_j)$ exceeds a fixed value independent of n. But the last disjunction is true for *every* $\tilde{P}_X$, since C_i and C_j are complementary events. We thus obtain the upper bound

$$-\frac{1}{n}\log Q_{XY}^n(\mathcal{A}_n) \;\leq\; \min_{\tilde{P}_X \in \mathcal{P}(\mathcal{X})} \{d(\tilde{P}_X, P_Y||Q) \;\vee\; d(\tilde{P}_X, \bar{P}_Y||Q)\} + \nu_n\;,$$

which, together with (4.5), yields

$$-\frac{1}{n}\log Q_{XY}^n(\mathcal{A}_n) \;\leq\; \theta^{(2)} + \nu_n\;.$$

Finally, by combining the bound for case 1 (equation (4.3)) with the above bound for case 2, we obtain the converse statement

$$\theta(2,3,\epsilon) \;\leq\; \theta^{(1)} \vee \theta^{(2)} \;. \qquad \triangle$$

For the system in which both encoders use two codewords, i.e., $M = N = 2$, we have the following result.

THEOREM 3.4. *Let* $\Pi = \{P_{XY}, \bar{P}_{XY}\}$, *where* $P_X \neq \bar{P}_X$ *and* $P_Y \neq \bar{P}_Y$. *If* $Q_{XY} > 0$, *then for* $0 < \epsilon < 1$,

$$\theta(2,2,\epsilon) \;=\; \theta^{(1)} \;\vee\; \theta^{(3)} \;,$$

where $\theta^{(1)}$ *is as defined in Theorem 3.3, and and* $\theta^{(3)}$ *is the supremum, over all partitions* $\{\Phi, \bar{\Phi}\}$ *of* $\mathcal{P}(\mathcal{X})$ *and* $\{\Psi, \bar{\Psi}\}$ *of* $\mathcal{P}(\mathcal{Y})$, *of the quantity*

$$d(\Phi \cup \{P_X\},\, \Psi \cup \{P_Y\}\,||Q) \;\wedge\; d(\bar{\Phi} \cup \{\bar{P}_X\},\, \bar{\Psi} \cup \{\bar{P}_Y\}\,||Q) \;. \qquad (4.6)$$

PROOF. *Direct part.* Since $M = 2$ as in the previous problem, we consider the same two candidates for the S_X encoder:

$$f: \quad C_1 \;=\; T_{X,\eta} \cup \bar{T}_{X,\eta} \;, \qquad C_2 \;=\; (T_{X,\eta} \cup \bar{T}_{X,\eta})^c$$

and

$$f': \quad C_1' \;=\; T_{X,\eta} \cup \bigcup_{\tilde{P}_X \in \Phi} \tilde{T}_X \;, \qquad C_2' \;=\; \bar{T}_{X,\eta} \cup \bigcup_{\tilde{P}_X \in \bar{\Phi}} \tilde{T}_X \;,$$

where $(\Phi, \bar{\Phi})$ form a partition of $\mathcal{P}(\mathcal{X}) - \mathcal{B}_\eta(P_X) - \mathcal{B}_\eta(\bar{P}_X)$. Observe that in this case $N = 2$ also, and thus it is no longer possible for the S_Y encoder to specify

whether y^n lies in $T_{Y,\eta}$, $\bar{T}_{Y,\eta}$ or $(T_{Y,\eta} \cup \bar{T}_{Y,\eta})^c$. Proceeding as for S_X, we propose the following two encoders for S_Y:

$$g: \quad F_1 \;=\; T_{Y,\eta} \cup \bar{T}_{Y,\eta} \;, \qquad F_2 \;=\; (T_{Y,\eta} \cup \bar{T}_{Y,\eta})^c$$

and

$$g': \quad F_1' \;=\; T_{Y,\eta} \cup \bigcup_{\tilde{P}_Y \in \Psi} \tilde{T}_Y \;, \qquad F_2' \;=\; \bar{T}_{Y,\eta} \cup \bigcup_{\tilde{P}_Y \in \bar{\Psi}} \tilde{T}_Y \;,$$

where $(\Psi, \bar{\Psi})$ are defined in a similar manner.

Given the above possibilities for encoding S_X and S_Y, there are only two reasonable choices for the acceptance region $\mathcal{A}_n$:

$$\mathcal{A}_n^{(1)} \;=\; C_1 \times F_1 \qquad \text{and} \qquad \mathcal{A}_n^{(3)} \;=\; (C_1' \times F_1') \cup (C_2' \times F_2') \;.$$

Note that the region $\mathcal{A}_n^{(1)}$ is identical to the one used in the proof of the previous theorem, whence we obtain

$$\theta(2,2,\epsilon) \;\geq\; \theta^{(1)} \;=\; d(\Pi_X, \Pi_Y || Q) \;.$$

To evaluate the error exponent associated with $\mathcal{A}_n^{(3)}$, we follow the corresponding procedure for $\mathcal{A}_n^{(2)}$ in the proof of Theorem 3.3. Since

$$\mathcal{A}_n^{(3)} \;=\; \Big(\bigcup_{\Phi \cup \mathcal{B}_\eta(P_X)} \tilde{T}_X \times \bigcup_{\Psi \cup \mathcal{B}_\eta(P_Y)} \tilde{T}_Y \Big) \cup \Big(\bigcup_{\bar{\Phi} \cup \mathcal{B}_\eta(\bar{P}_X)} \tilde{T}_X \times \bigcup_{\bar{\Psi} \cup \mathcal{B}_\eta(\bar{P}_Y)} \tilde{T}_Y \Big) \;,$$

we obtain

$$-\lim_n \frac{1}{n} \log Q^n(\mathcal{A}_n^{(3)}) \;=\; d(\Phi \cup \{P_X\},\, \Psi \cup \{P_Y\} \,||Q) \wedge d(\bar{\Phi} \cup \{\bar{P}_X\},\, \bar{\Psi} \cup \{\bar{P}_Y\} \,||Q) \;.$$

Once again, it is legitimate to assume that in the above equation, $\{\Phi, \bar{\Phi}\}$, $\{\Psi, \bar{\Psi}\}$ constitute partitions of the entire spaces $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$, respectively. The best

error exponent attainable by a sequence of acceptance regions of the form $\mathcal{A}_n^{(3)}$ is therefore

$$\theta^{(3)} \;=\; \sup_{\Phi,\Psi} \{ d(\Phi \cup \{P_X\},\, \Psi \cup \{P_Y\} \,||Q) \;\wedge\; d(\bar{\Phi} \cup \{\bar{P}_X\},\, \bar{\Psi} \cup \{\bar{P}_Y\} \,||Q) \} \;.$$

We conclude that

$$\theta(2,2,\epsilon) \;\geq\; \theta^{(1)} \;\vee\; \theta^{(3)} \;.$$

Converse part. In this case every admissible acceptance region $\mathcal{A}_n$ can be written as

$$\mathcal{A}_n \;=\; (C_1 \times F_1) \;\cup\; (C_2 \times F_2),$$

where $C_2 = C_1^c$, while F_1, F_2 are constrained by $F_2 \in \{\emptyset, \mathcal{Y}^n, F_1^c\}$. As in the proof of Theorem 3.3, two cases may arise.

Case 1. For i and j distinct, we have

$$P_{XY}^n(C_i \times F_i) \;\geq\; P_{XY}^n(C_j \times F_j) \qquad \text{and} \qquad \bar{P}_{XY}^n(C_i \times F_i) \;\geq\; \bar{P}_{XY}^n(C_j \times F_j) \;.$$

This is same as Case 1 in the proof of Theorem 3.3, whence we obtain

$$-\frac{1}{n} \log Q^n(\mathcal{A}_n) \;\leq\; \theta^{(1)} + \nu_n \;.$$

Note that this case subsumes the situation in which F_2 is empty.

Case 2. For i and j distinct, we have

$$P_{XY}^n(C_i \times F_i) \;\geq\; P_{XY}^n(C_j \times F_j) \qquad \text{and} \qquad \bar{P}_{XY}^n(C_i \times F_i) \;<\; \bar{P}_{XY}^n(C_j \times F_j) \;.$$

We easily deduce that

$$P_X^n(C_i) \;\geq\; \frac{1-\epsilon}{2} \;, \qquad\qquad P_Y^n(F_i) \;\geq\; \frac{1-\epsilon}{2} \;,$$

and

$$\bar{P}_X^n(C_j) \geq \frac{1-\epsilon}{2}, \qquad \bar{P}_Y^n(F_j) \geq \frac{1-\epsilon}{2}.$$

Let us define the classes

$$\Phi_n = \{\tilde{P}_X : \tilde{P}_X^n(C_i) \geq \frac{1}{2}\}, \qquad \Psi_n = \{\tilde{P}_Y : \tilde{P}_Y^n(F_i) \geq \frac{1}{2}\},$$

and

$$\Phi_n^* = \{\tilde{P}_X : \tilde{P}_X^n(C_j) > \frac{1}{2}\}, \qquad \Psi_n^* = \{\tilde{P}_Y : \tilde{P}_Y^n(F_j) > \frac{1}{2}\}.$$

Since C_1 and C_2 are complementary, $\Phi_n^* = \Phi_n^c$. For F_1 and F_2, we have either $F_2 = F_1^c$ or $F_2 = \mathcal{Y}^n$. In the former case we have again $\Psi_n^* = \Psi_n^c$, while in the latter, either Ψ_n or Ψ_n^* is equal to $\mathcal{P}(\mathcal{Y})$.

By the foregoing discussion, all marginal distributions $\tilde{P}_X \in \Phi_n \cup \{P_X\}$, $\tilde{P}_Y \in \Psi_n \cup \{P_Y\}$, satisfy

$$\tilde{P}_X^n(C_i) \geq \frac{1-\epsilon}{2} \qquad \text{and} \qquad \tilde{P}_Y^n(F_i) \geq \frac{1-\epsilon}{2}.$$

Applying Theorem 2.3, we obtain

$$-\frac{1}{n}\log Q_{XY}^n(C_i \times F_i) \leq d(\Phi_n \cup \{P_X\}, \Psi_n \cup \{P_Y\} \,||Q) + \nu_n. \tag{4.7}$$

Similarly for $C_j \times F_j$ we have

$$-\frac{1}{n}\log Q_{XY}^n(C_j \times F_j) \leq d(\Phi_n^* \cup \{\bar{P}_X\}, \Psi_n^* \cup \{\bar{P}_Y\} \,||Q) + \nu_n. \tag{4.8}$$

We must show that the smaller of the two bounds appearing in equations (4.7) and (4.8) is less than or equal to $\theta^{(3)}$ as defined in the statement of the theorem. This is certainly true if $\Psi_n^* = \Psi_n^c$, since we can then take

$$\{\Phi, \bar{\Phi}\} = \{\Phi_n, \Phi_n^*\} \qquad \text{and} \qquad \{\Psi, \bar{\Psi}\} = \{\Psi_n, \Psi_n^*\}$$

in the definition of $\theta^{(3)}$. Otherwise, if w.l.o.g. $\Psi_n^* = \mathcal{P}(\mathcal{Y})$, the same conclusion can be reached by taking

$$\{\Phi, \bar{\Phi}\} = \{\Phi_n, \Phi_n^*\} \qquad \text{and} \qquad \{\Psi, \bar{\Psi}\} = \{\Psi_n, \Psi_n^c\} \ .$$

Thus we have obtained

$$-\frac{1}{n} \log Q_{XY}^n(\mathcal{A}_n) \ \leq \ \theta^{(3)} + \nu_n \ .$$

This, together with our result for Case 1, yields the converse statement

$$\theta(2,2,\epsilon) \ \leq \ \theta^{(1)} \vee \theta^{(3)} \ . \qquad \triangle$$

REMARKS. (a) It is shown in the Appendix A that $\theta^{(3)}$ can also expressed in the simpler form

$$\theta^{(3)} \ = \ \{D(P_X||Q_X) \ \wedge \ D(\bar{P}_Y||Q_Y)\} \ \vee \ \{D(\bar{P}_X||Q_X) \ \wedge \ D(P_Y||Q_Y)\} \ . \quad (4.9)$$

This characterization also simplifies the determination of the maximizing classes Φ and Ψ in the original definition of $\theta^{(3)}$.

(b) The definition of the asymptotically optimal acceptance regions $\mathcal{A}_n^{(2)}$ and $\mathcal{A}_n^{(3)}$ in the proofs of Theorems 3.3 and 3.4 depends implicitly on the alternative distribution Q_{XY} through the choice of the optimal classes Φ and Ψ; this is not the case with $\mathcal{A}_n^{(1)}$.

(c) Which of the alternative exponents is the dominant one depends on Π and Q_{XY}. To show this, in what follows we let Q_{XY} be a product distribution on $\mathcal{X} \times \mathcal{Y}$, i.e., $Q_{XY} = Q_X \times Q_Y$, where $Q_X > 0$, $Q_Y > 0$. Then it is quite

straightforward to show that

$$\theta^{(1)} = \{D(P_X||Q_X) \wedge D(\bar{P}_X||Q_X)\} + \{D(P_Y||Q_Y) \wedge D(\bar{P}_Y||Q_Y)\} ;$$

$$\theta^{(2)} = \{(D(P_X||Q_X) + D(P_Y||Q_Y)) \wedge D(\bar{P}_Y||Q_Y)\}$$
$$\vee \{D(P_Y||Q_Y) \wedge (D(\bar{P}_X||Q_X) + D(\bar{P}_Y||Q_Y))\} ;$$

$$\theta^{(3)} = \{D(P_X||Q_X) \wedge D(\bar{P}_Y||Q_Y)\} \vee \{D(\bar{P}_X||Q_X) \wedge D(P_Y||Q_Y)\} .$$

Consider first the case in which $\bar{P}_X = Q_X$ and $D(\bar{P}_Y||Q_Y) > D(P_X||Q_X) + D(P_Y||Q_Y)$. From the above we obtain

$$\theta^{(1)} = D(P_Y||Q_Y) , \quad \theta^{(2)} = D(P_X||Q_X) + D(P_Y||Q_Y) , \quad \theta^{(3)} = D(P_X||Q_X) .$$

Thus provided all above divergences are positive and distinct, we obtain either $\theta^{(2)} > \theta^{(3)} > \theta^{(1)}$ or $\theta^{(2)} > \theta^{(1)} > \theta^{(3)}$.

As another example, consider the situation in which all distributions are distinct, and

$$D(P_X||Q_X) = D(\bar{P}_X||Q_X) , \qquad D(P_Y||Q_Y) = D(\bar{P}_Y||Q_Y) .$$

Then

$$\theta^{(1)} = D(P_X||Q_X) + D(P_Y||Q_Y) , \qquad \theta^{(2)} = D(P_Y||Q_Y) ,$$

$$\theta^{(3)} = D(P_X||Q_X) \wedge D(P_Y||Q_Y) .$$

We thus obtain either $\theta^{(1)} > \theta^{(2)} > \theta^{(3)}$ or $\theta^{(1)} > \theta^{(2)} = \theta^{(3)}$.

3.5. Dependence of the error exponent on ϵ.

Theorems 3.3 and 3.4 were derived under the assumption that the distributions P_{XY} and $\bar{P}_{XY}$ have distinct X and Y marginals. As it turns out, the

conclusions of these theorems are true even if this assumption is not. Indeed, it is easy to show that if $P_X = \bar{P}_X$ or $P_Y = \bar{P}_Y$, then $\mathcal{A}_n^{(1)}$ is optimal, and $\theta^{(1)}$ dominates both $\theta^{(2)}$ and $\theta^{(3)}$.

If $|\Pi| > 2$, and the codebook sizes M and N are fixed at levels below the thresholds given in Theorem 3.2, then it is still possible to derive versions of Theorems 3.3 and 3.4 in which the acceptance regions $\mathcal{A}_n^{(2)}$ and $\mathcal{A}_n^{(3)}$ are constructed by first grouping distributions in Π_X and Π_Y together, and then partitioning $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ appropriately. Our final result in this chapter illustrates this procedure, and more importantly, it reveals a hitherto unseen aspect of this problem: specifically, if the marginals of some distributions in Π coincide, the error exponent may depend on the level ϵ. This is certainly a surprising discovery, considering the chain of strong converse theorems which have been derived in [10-12], and in this work.

NOTATION. 1_X denotes the set of degenerate distributions on $\mathcal{P}(\mathcal{X})$.

THEOREM 3.5. *Let* $\Pi < \infty$, $M = 2$, *and* $N \geq |\Pi_Y| + 1$. *Also, let* $\{\Delta, \bar{\Delta}\}$ *denote a partition of* Π. *If* $Q_{XY} > 0$, *then for* $\epsilon \in (0, 1/2) \cup (1/2, 1)$, *the following is true:*

$$\theta(2, N, \epsilon) \;=\; \theta^{(1)} \;\vee\; \theta^{(4)}(\epsilon),$$

where

$$\theta^{(1)} \;=\; d(\Pi_X, \Pi_Y || Q)\;,$$

$$\theta^{(4)}(\epsilon) \;=\; \begin{cases} \max_{\Delta, \bar{\Delta}:\ \Delta_X \cap \bar{\Delta}_X = \emptyset} \tau(\Delta, \bar{\Delta}), & \text{if } 0 < \epsilon < \frac{1}{2}; \\ \max_{\substack{\Delta, \bar{\Delta}: \\ \Delta_X \cap \bar{\Delta}_X \cap 1_X = \emptyset}} \tau(\Delta, \bar{\Delta}), & \text{if } \frac{1}{2} < \epsilon < 1, \end{cases}$$

and

$$\tau(\Delta,\bar{\Delta}) = d(\Delta_X,\Delta_Y||Q) \ \wedge \ d(\bar{\Delta}_X,\bar{\Delta}_Y||Q)$$
$$\wedge \inf_{\tilde{P}_X}\{d(\tilde{P}_X,\Delta_Y||Q) \ \vee \ d(\tilde{P}_X,\bar{\Delta}_Y||Q)\} \ .$$

REMARK. We have been unable to evaluate $\theta(2,N,1/2)$.

PROOF. *Direct part.* Once again it is feasible to construct $\mathcal{A}_n^{(1)}$ as defined in the proof of Theorem 3.3, whence we obtain $\theta(2,N,\epsilon) \geq \theta^{(1)}$.

To construct $\mathcal{A}_n^{(2)}$ by analogy to Theorem 3.3, we partition the space Π_X into Λ, $\bar{\Lambda}$, and the space $\mathcal{P}(\mathcal{X}) - \mathcal{B}_\eta(\Pi_X)$ into Φ, $\bar{\Phi}$. We then have

$$\mathcal{A}_n^{(2)} =$$
$$\Big(\bigcup_{\tilde{P}_X \in \Phi \cup \mathcal{B}_\eta(\Lambda)} \tilde{T}_X \times \bigcup_{\tilde{P}_{XY} \in \Pi:\ \tilde{P}_X \in \Lambda} \tilde{T}_{Y,\eta} \Big) \cup \Big(\bigcup_{\tilde{P}_X \in \bar{\Phi} \cup \mathcal{B}_\eta(\bar{\Lambda})} \tilde{T}_X \times \bigcup_{\tilde{P}_{XY} \in \Pi:\ \tilde{P}_X \in \bar{\Lambda}} \tilde{T}_{Y,\eta} \Big)$$

which is readily seen to satisfy the type I error constraint for every ϵ and every distribution in Π.

Note that instead of partitioning Π_X into Λ and $\bar{\Lambda}$, one can begin by partitioning Π itself into Δ and $\bar{\Delta}$ such that $\Delta_X \cap \bar{\Delta}_X = \emptyset$. Then one can write equivalently

$$\mathcal{A}_n^{(2)} = \Big(\bigcup_{\tilde{P}_X \in \Phi \cup \mathcal{B}_\eta(\Delta_X)} \tilde{T}_X \times \bigcup_{\tilde{P}_Y \in \Delta_Y} \tilde{T}_{Y,\eta} \Big) \cup \Big(\bigcup_{\tilde{P}_X \in \bar{\Phi} \cup \mathcal{B}_\eta(\bar{\Delta}_X)} \tilde{T}_X \times \bigcup_{\tilde{P}_Y \in \bar{\Delta}_Y} \tilde{T}_{Y,\eta} \Big) \ ,$$

and by the argument given in the proof of Theorem 3.3,

$$\theta(2,N,\epsilon) \ \geq \ \tau(\Delta,\bar{\Delta}) = d(\Delta_X,\Delta_Y||Q) \ \wedge \ d(\bar{\Delta}_X,\bar{\Delta}_Y||Q)$$
$$\wedge \inf_{\tilde{P}_X}\{d(\tilde{P}_X,\Delta_Y||Q) \ \vee \ d(\tilde{P}_X,\bar{\Delta}_Y||Q)\} \ .$$

Taking the maximum over all partitions $\{\Delta, \bar{\Delta}\}$ of Π satisfying $\Delta_X \cap \bar{\Delta}_X = \emptyset$, we obtain for all $\epsilon \in (0,1)$,

$$\theta(2, N, \epsilon) \;\geq \max_{\Delta, \bar{\Delta}:\ \Delta_X \cap \bar{\Delta}_X = \emptyset} \tau(\Delta, \bar{\Delta}) \ .$$

The constraint $\Delta_X \cap \bar{\Delta}_X = \emptyset$ is essential in the above construction of $\mathcal{A}_n^{(2)}$; its removal would allow

$$C_1' \;=\; \bigcup_{\tilde{P}_X \in \Phi \cup \mathcal{B}_\eta(\Delta_X)} \tilde{T}_X \qquad \text{and} \qquad C_2' \;=\; \bigcup_{\tilde{P}_X \in \Phi \cup \mathcal{B}_\eta(\bar{\Delta}_X)} \tilde{T}_X$$

to have nonempty intersection and hence be inadmissible under the given compression scheme. If, however, $1/2 < \epsilon < 1$, then it is possible to relax the said constraint to

$$\Delta_X \;\cap\; \bar{\Delta}_X \;\cap\; 1_X \;=\; \emptyset$$

in the following manner. For every $\tilde{P}_X$ that lies in $\mathcal{B}_\eta(\Delta_X \cap \bar{\Delta}_X)$ (and hence not in 1_X if η is properly chosen), we can partition $\tilde{T}_X$ into two sets $\tilde{T}_X^+$ and $\tilde{T}_X^-$ of sizes that differ by at most 1, and redefine C_1' and C_2' by

$$C_1' \;=\; \bigcup_{\tilde{P}_X \in \Phi \,\cup \mathcal{B}_\eta(\Delta_X - \bar{\Delta}_X)} \tilde{T}_X \cup \bigcup_{\tilde{P}_X \in \mathcal{B}_\eta(\Delta_X \cap \bar{\Delta}_X)} \tilde{T}_X^+$$

and

$$C_2' \;=\; \bigcup_{\tilde{P}_X \in \bar{\Phi} \,\cup \mathcal{B}_\eta(\bar{\Delta}_X - \Delta_X)} \tilde{T}_X \cup \bigcup_{\tilde{P}_X \in \mathcal{B}_\eta(\Delta_X \cap \bar{\Delta}_X)} \tilde{T}_X^- \ .$$

We can then complete the construction of $\mathcal{A}_n^{(2)}$ in the usual manner.

It is easily seen that for every $P_{XY} \in \Pi$ such that $P_X \notin \Delta_X \cap \bar{\Delta}_X$, and every $\epsilon \in (0,1)$,

$$P_{XY}^n(\mathcal{A}_n^{(2)}) \;\geq\; 1 - \epsilon$$

for n sufficiently large. The same is true for for every $P_{XY} \in \Pi$ such that $P_X \in \Delta_X \cap \bar{\Delta}_X$, if $\epsilon \in (1/2, 1)$. To see this, let w.l.o.g. $P_{XY} \in \Delta$. Then

$$\begin{aligned} P_{XY}^n(\mathcal{A}_n^{(2)}) &\geq P_{XY}^n\Big(\bigcup_{\tilde{P}_X \in \mathcal{B}_\eta(P_X)} \tilde{T}_X^+ \times T_{Y,\eta}\Big) \\ &\geq P_X^n\Big(\bigcup_{\tilde{P}_X \in \mathcal{B}_\eta(P_X)} \tilde{T}_X^+\Big) + P_Y^n(T_{Y,\eta}) - 1 \\ &\geq \frac{1}{2} - \lambda_n + 1 - \frac{|\mathcal{Y}|}{4n\eta^2} - 1 \end{aligned}$$

where $\lambda_n \to 0$ since $\mathcal{B}_\eta(P_X)$ contains no degenerate distributions. We conclude that for n sufficiently large,

$$P_{XY}^n(\mathcal{A}_n^{(2)}) \geq 1 - \epsilon \, .$$

By computing the error exponent as before, we obtain for $1/2 < \epsilon < 1$,

$$\theta(2, N, \epsilon) \geq \max_{\substack{\Delta, \bar{\Delta}: \\ \Delta_X \cap \bar{\Delta}_X \cap \iota_X = \emptyset}} \tau(\Delta, \bar{\Delta}) \, .$$

This concludes the proof of the positive part.

Converse part. As in the proof of the converse part of Theorem 3.3, we express $\mathcal{A}_n$ as

$$\mathcal{A}_n = (C_1 \times F_1) \cup (C_2 \times F_2),$$

where C_1 and C_2 form a partition of $\mathcal{X}^n$, and at most one of F_1, F_2 may be empty. Once again, two cases may arise.

Case 1. For i and j distinct, we have

$$(\forall P_{XY} \in \Pi) \qquad P_{XY}^n(C_i \times F_i) \geq P_{XY}^n(C_j \times F_j) \, .$$

This implies that

$$-\frac{1}{n} \log Q^n(\mathcal{A}_n) \leq \theta^{(1)} + \nu_n \, .$$

Case 2. The sets Δ and $\bar{\Delta}$ defined below form a nontrivial partition of Π

$$\Delta = \{P_{XY} \in \Pi\colon\ P_{XY}^n(C_1 \times F_1) \geq P_{XY}^n(C_2 \times F_2)\}\ ,$$

$$\bar{\Delta} = \{P_{XY} \in \Pi\colon\ P_{XY}^n(C_1 \times F_1) < P_{XY}^n(C_2 \times F_2)\}\ .$$

We claim further that $\Delta_X \cap \bar{\Delta}_X \cap 1_X = \emptyset$. Indeed, if there exist $P_{XY} \in \Delta$ and $\tilde{P}_{XY} \in \bar{\Delta}$ such that $P_X = \tilde{P}_X$, then

$$P_X^n(C_1)\ \geq\ \frac{1-\epsilon}{2}, \qquad \tilde{P}_X^n(C_2)\ =\ P_X^n(C_2)\ >\ \frac{1-\epsilon}{2}\ .$$

Since C_1 and C_2 are complementary and have positive probability under P_X^n, P_X cannot be degenerate.

As in Case 2 in the proof of the converse of Theorem 3.3, we obtain for all $\epsilon \in (0,1)$,

$$-\frac{1}{n}\log Q^n(\mathcal{A}_n)\ \leq\ \tau(\Delta,\bar{\Delta}) + \nu_n\ .$$

It remains to show that if $\epsilon \in (0,1/2)$, the above bound is also valid for a partition $\{\Omega,\bar{\Omega}\}$ of Π such that $\Omega_X \cap \bar{\Omega}_X = \emptyset$. To construct such a partition, we argue as follows.

For $P_X \in \Pi_X$, we consider the set $\mathcal{H}(P_X)$ of distributions in Π that have P_X as X-marginal:

$$\mathcal{H}(P_X) \stackrel{\text{def}}{=} \{\tilde{P}_{XY} \in \Pi\colon \tilde{P}_X = P_X\}\ .$$

We let $\lambda > 0$ be independent of n, and we assume for the moment that for every $P_X \in \Pi_X$, we can find $i \in \{1,2\}$ such that

$$(\forall \tilde{P}_{XY} \in \mathcal{H}(P_X)) \qquad \tilde{P}_{XY}^n(C_i \times F_i) \geq \lambda\ . \tag{5.1}$$

If so, then we can partition Π_X into Λ_1 and Λ_2 by placing each of the members P_X of Π_X in Λ_i iff i is the smallest index for which the above relationship holds. This in turn yields a partition $\Omega, \bar{\Omega}$ of Π through

$$\Omega \;=\; \bigcup_{P_X \in \Lambda_1} \mathcal{H}(P_X) \qquad \text{and} \qquad \bar{\Omega} \;=\; \bigcup_{P_X \in \Lambda_2} \mathcal{H}(P_X)\;.$$

Clearly $\Omega_X = \Lambda_1$, $\bar{\Omega}_X = \Lambda_2$, and from the definition of Λ_i and relationship (5.1), we obtain the desired bound

$$-\frac{1}{n}\log Q^n(\mathcal{A}_n) \;\leq\; \tau(\Omega, \bar{\Omega}) + \nu_n\;.$$

Thus the issue is to prove that for suitable $\lambda > 0$, every $P_X \in \Pi_X$ is such that (5.1) holds for $i = 1$ or $i = 2$. By definition of the classes Δ and $\bar{\Delta}$, this is true for $P_X \in \Delta_X - \bar{\Delta}_X$ and $P_X \in \bar{\Delta}_X - \Delta_X$. To show that it is also true for $P_X \in \Delta_X \cap \bar{\Delta}_X$, assume the contrary, namely that there exists $P_{XY} \in \Delta$ and $\tilde{P}_{XY} \in \bar{\Delta}$ with $\tilde{P}_X = P_X$ and

$$P^n_{XY}(C_2 \times F_2) < \lambda, \qquad \tilde{P}^n_{XY}(C_1 \times F_1) < \lambda\;.$$

This implies that

$$P^n_X(C_1) \;\geq\; P^n_{XY}(C_1 \times F_1) \;>\; 1 - \epsilon - \lambda\;,$$

$$P^n_X(C_2) \;\geq\; \tilde{P}^n_{XY}(C_2 \times F_2) \;>\; 1 - \epsilon - \lambda\;,$$

and hence

$$P^n_X(C_1) + P^n_X(C_2) \;>\; 2 - 2\epsilon - 2\lambda\;.$$

Thus if $\epsilon < 1/2$, we can set $\lambda = (1-2\epsilon)/3 > 0$ to obtain the desired contradiction:

$$P^n_X(C_1) + P^n_X(C_2) \;>\; 1 + \lambda\;. \qquad \triangle$$

It should be emphasized that the dependence of $\theta(2, N, \epsilon)$ on ϵ is nontrivial. If we consider the simple setup in which $\Pi = \{P_{XY}, \bar{P}_{XY}, \tilde{P}_{XY}\}$ with $P_X = \tilde{P}_X$ and $\bar{P}_Y = \tilde{P}_Y$, then it is possible to choose the above distributions so that the error exponent for $0 < \epsilon < 1/2$ is strictly less than for $1/2 < \epsilon < 1$. This is demonstrated in the following example:

EXAMPLE 3.1. We define for simplicity

$$d_{\tilde{X}\tilde{Y}} \stackrel{\text{def}}{=} d(\tilde{P}_X, \bar{P}_Y || Q).$$

Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$; $\Pi = \{P_{XY}, \bar{P}_{XY}, \tilde{P}_{XY}\}$, where

$$P_X = \tilde{P}_X = (0.4,\ 0.6)^T,$$

$$\bar{P}_X = (0.54,\ 0.46)^T,$$

$$P_Y = (0.35,\ 0.65)^T,$$

$$\bar{P}_Y = \tilde{P}_Y = (0.01,\ 0.99)^T;$$

and

$$Q_{XY} = \begin{pmatrix} 0.1 & 0.4 \\ 0.05 & 0.45 \end{pmatrix}.$$

We thus have $\Pi_X = \{P_X, \bar{P}_X\}$ and $\Pi_Y = \{P_Y, \bar{P}_Y\}$. Using the above Theorem, we have

$$\begin{aligned} \theta^{(1)} &= \min_{\hat{P}_X \in \Pi_X,\ \hat{P}_Y \in \Pi_Y} d_{\hat{X}\hat{Y}} \\ &= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}Y} \wedge d_{\bar{X}\bar{Y}} \\ &= 0.235 \wedge 0.194 \wedge 0.176 \wedge 0.192 \\ &= 0.176 \end{aligned}$$

The only nontrivial partition $\{\Delta, \bar{\Delta}\}$ of Π such that $\Delta_X \cap \bar{\Delta}_X = \emptyset$ is given by $\Delta = \{P_{XY}, \tilde{P}_{XY}\}$, $\bar{\Delta} = \{\bar{P}_{XY}\}$. This yields $\Delta_X = \{P_X\}$, $\bar{\Delta}_X = \{\bar{P}_X\}$ and $\Delta_Y = \{P_Y, \bar{P}_Y\}$, $\bar{\Delta}_Y = \{\bar{P}_Y\}$. Hence, for $0 < \epsilon < 0.5$,

$$\begin{aligned}
\theta^{(4)} &= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}\bar{Y}} \wedge \min_{\hat{P}_X}\{(d_{\hat{X}Y} \wedge d_{\hat{X}\bar{Y}}) \vee d_{\hat{X}\bar{Y}}\} \\
&= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}\bar{Y}} \wedge \min_{\hat{P}_X} d_{\hat{X}\bar{Y}} \\
&= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}\bar{Y}} \wedge D(\bar{P}_Y||Q_Y) \\
&= d_{XY} \wedge D(\bar{P}_Y||Q_Y) \\
&= 0.235 \wedge 0.179 \\
&= 0.179
\end{aligned}$$

This implies that $\theta(2, N, \epsilon) = 0.179$ if $0 < \epsilon < 0.5$. The other two nontrivial partitions of Π are $\{\{P_{XY}, \bar{P}_{XY}\}, \{\tilde{P}_{XY}\}\}$ and $\{\{P_{XY}\}, \{\bar{P}_{XY}, \tilde{P}_{XY}\}\}$ with corresponding projections $\{\{P_X, \bar{P}_X\}, \{P_X\}\}$, $\{\{P_Y, \bar{P}_Y\}, \{\bar{P}_Y\}\}$ for the first partition, and $\{\{P_X\}, \{P_X, \bar{P}_X\}\}$, $\{\{P_Y\}, \{\bar{P}_Y\}\}$ for the second. The corresponding error exponents are

$$\begin{aligned}
\tau_1(\Delta, \bar{\Delta}) &= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}\bar{Y}} \wedge d_{\bar{X}Y} \wedge \min_{\hat{P}_X}\{(d_{\hat{X}Y} \wedge d_{\hat{X}\bar{Y}}) \vee d_{\hat{X}\bar{Y}}\} \\
&= d_{XY} \wedge d_{\bar{X}Y} \wedge D(\bar{P}_Y||Q_Y) \\
&= 0.235 \wedge 0.176 \wedge 0.179 \\
&= 0.176
\end{aligned}$$

$$\begin{aligned}
\tau_2(\Delta, \bar{\Delta}) &= d_{XY} \wedge d_{X\bar{Y}} \wedge d_{\bar{X}\bar{Y}} \wedge \min_{\hat{P}_X}\{d_{\hat{X}Y} \vee d_{\hat{X}\bar{Y}}\} \\
&= 0.235 \wedge 0.194 \wedge 0.192 \wedge 0.181 \\
&= 0.181
\end{aligned}$$

This implies that $\theta(2, N, \epsilon) = 0.181$ if $0.5 < \epsilon < 1$. Hence

$$\theta(2, N, \epsilon) \;=\; \begin{cases} 0.179, & \text{if } 0 < \epsilon < \frac{1}{2}\ ; \\ 0.181, & \text{if } \frac{1}{2} < \epsilon < 1\ . \end{cases}$$

3.6. Concluding remarks

We have been unable to evaluate $\theta(2, N, 1/2)$ in Theorem 3.5. However from the proof of that theorem, we can bound $\theta(2, N, 1/2)$ as follows:

$$\theta^{(1)} \vee \max_{\Delta, \bar{\Delta}:\ \Delta_X \cap \bar{\Delta}_X = \emptyset} \tau(\Delta, \bar{\Delta}) \;\leq\; \theta(2, N, 1/2) \;\leq\; \theta^{(1)} \vee \max_{\substack{\Delta, \bar{\Delta}: \\ \Delta_X \cap \bar{\Delta}_X \cap 1_X = \emptyset}} \tau(\Delta, \bar{\Delta})\ .$$

If $P_X \neq \bar{P}_X$ for all $P_{XY}, \bar{P}_{XY} \in \Pi$, then $\theta(2, N, \epsilon)$ is independent of ϵ and the above inequalities become equalities. This shows that Theorem 3.3 is actually subsumed under Theorem 3.5.

Theorem 3.5 can also be extended in a straightforward manner to the case in which $2 \leq M \leq |\Pi_X|$, for values of ϵ in the range $(0, \frac{1}{M}) \cup (\frac{M-1}{M}, 1)$.

CHAPTER 4

MULTITERMINAL HYPOTHESIS TESTING WITH FIXED-RATE DATA COMPRESSION

4.1. Introduction.

In the previous two chapters, we obtained closed-form characterizations of optimal error exponents for multiterminal hypothesis testing under *asymptotically zero-rate* compression. In this chapter we discuss the problem of finding similar characterizations in the framework of multiterminal hypothesis testing under *fixed-rate* compression.

This problem was first studied by Ahlswede and Csiszár [10], where an achievable lower bound on the error exponent for hypothesis testing with one-sided compression was obtained. In the same work, it was shown by an example that this bound was not tight. Han [11] was able to derive better lower bounds but the question of their tightness remained open. In this chapter we show by example that one of the two bounds derived by Han is not tight. We also propose a sequence of lower bounds on the error exponent that is asymptotically tight, but do not succeed in providing a single-letter characterization for that exponent (single-letter characterizations, such as those given in Chapters 2 and 3, are expressible in terms of finitely many random variables taking values in finite sets [17]).

4.2. Problem statement and preliminaries.

(a) General notation. The same notation and problem definition introduced in Section 2.2.(a) will also be used here except for the code rate constraints,

which should be modified to

$$\frac{1}{n}\log M_n < R_X \tag{2.1}$$

and

$$\frac{1}{n}\log N_n < R_Y \ , \tag{2.2}$$

where $R_X,\ R_Y > 0$ are the rate constraints on S_X and S_Y encoders, respectively.

(b) Simple hypothesis testing with fixed rate. The optimal acceptance region for testing $H_0 : P$ versus $H_1 : Q$ at a given level $\epsilon \in (0,1)$ is one that minimizes $Q^n_{XY}(\mathcal{A}_n)$ over all acceptance regions $\mathcal{A}_n$ that:

(C1) yield a value of $P^n_{XY}(\mathcal{A}^c_n)$ less than or equal to ϵ;

and

(C2) satisfy the appropriate compression constraints:

- (2.1) and ((2.3) of Chapter 2) for one-sided compression of X^n;
- (2.2) and ((2.4) of Chapter 2) for one-sided compression of Y^n;
- (2.1), (2.2), and ((2.3) and (2.4) of Chapter 2) for two-sided compression.

The resulting minimum probability of type II error for two-sided compression is denoted by $\beta_n(R_X, R_Y, \epsilon)$, and the associated error exponent is given by

$$\theta(R_X, R_Y, \epsilon) \stackrel{\text{def}}{=} -\lim_n \frac{1}{n}\log \beta_n(R_X, R_Y, \epsilon) \ ,$$

provided the limit on the right-hand side exists.

For one-sided compression of X^n the corresponding minimum type II error rate is denoted by $\beta_n(R_X, \epsilon)$, and the error exponent is given by

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

Similar notation is used for one-sided compression of Y^n.

In this chapter we will restrict ourselves to one-sided compression of X^n and hence use R instead of R_X.

(c) Conditional typical sequences. We summarize here some of the basic definitions and relationships on conditional typical sequences.

The conditional type of a sequence $y^n \in \mathcal{Y}^n$ given $x^n \in \mathcal{X}^n$ is the distribution $\lambda_{y|x}$ on $\mathcal{X} \times \mathcal{Y}$ defined by the relationship

$$(\forall (a,b) \in \mathcal{X} \times \mathcal{Y}) \qquad \lambda_{y|x}(b|a) \stackrel{\text{def}}{=} \lambda_{xy}(a,b)/\lambda_x(a), \qquad \text{if } \lambda_x(a) > 0 \ .$$

For any $x^n \in \mathcal{X}^n$, let $\mathcal{P}_n(\mathcal{Y}|x^n)$ denote the set of all conditional types given x^n:

$$\mathcal{P}_n(\mathcal{Y}|x^n) \stackrel{\text{def}}{=} \{\hat{P}_{Y|X} : \hat{P}_{Y|X}\lambda_x \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})\} \ .$$

Given $x^n \in \mathcal{X}^n$ and $\hat{P}_{Y|X} \in \mathcal{P}_n(\mathcal{Y}|x^n)$, the set of sequences in $\mathcal{Y}^n$ that have conditional type $\hat{P}_{Y|X}$ given x^n is defined by

$$\hat{T}_{Y|X}(x^n) \stackrel{\text{def}}{=} \{y^n \in \mathcal{Y}^n : (\forall a \in \mathcal{X},\ b \in \mathcal{Y}) \quad \lambda_{xy}(a,b) = \hat{P}_{Y|X}(b|a)\lambda_x(a)\} \ .$$

For arbitrary stochastic matrices $P_{Y|X}$, $Q_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$ and a distribution P_X on $\mathcal{X}$, the conditional entropy and conditional divergence functionals are defined by

$$H(Y|X) \stackrel{\text{def}}{=} -\sum_{x,y} P_{XY}(x,y) \log P_{Y|X}(y|x) \ ,$$

$$D(P_{Y|X}||Q_{Y|X}|P_X) \stackrel{\text{def}}{=} \sum_{x,y} P_{XY}(x,y) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \ .$$

Moreover, for $\eta > 0$, we will denote by $T_{Y|X,\eta}(x^n)$ the set of sequences $y^n \in \mathcal{Y}^n$ that are $(P_{Y|X}, \eta)$-typical under x^n:

$$T_{Y|X,\eta}(x^n) \stackrel{\text{def}}{=} \{y^n \in \mathcal{Y}^n : ||\lambda_{xy} - P_{Y|X}\lambda_x|| \le \eta,\ \lambda_{y|x} \ll P_{Y|X}\}$$

$$= \bigcup_{\substack{\hat{P}_{Y|X} \in \mathcal{P}_n(\mathcal{Y}|x^n): \\ ||\hat{P}_{Y|X}\lambda_x - P_{Y|X}\lambda_x|| \le \eta,\ \hat{P}_{Y|X} \ll P_{Y|X}}} \hat{T}_{Y|X}(x^n) \ .$$

LEMMA 4.1. *For any $\hat{P}_X \in \mathcal{P}_n(\mathcal{X})$, $x^n \in \hat{T}_X$, and $\hat{P}_{Y|X} \in \mathcal{P}_n(\mathcal{Y}|x^n)$,*

$$\frac{1}{|\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})|} \exp[-nH(\hat{Y}|\hat{X})] \le |\hat{T}^n_{Y|X}(x^n)| \le \exp[-nH(\hat{Y}|\hat{X})] \ ,$$

$$\frac{1}{|\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})|} \exp[-nD(\hat{P}_{Y|X}||Q_{Y|X}|\hat{P}_X)] \le Q^n_{Y|X}(\hat{T}_{Y|X}(x^n)|x^n)$$
$$\le \exp[-nD(\hat{P}_{Y|X}||Q_{Y|X}|\hat{P}_X)] \ .$$

In addition, if $P_{Y|X}$ is any stochastic matrix, then

$$P^n_{Y|X}(T_{Y|X,\eta}(x^n)|x^n) \ \ge \ 1 - \frac{|\mathcal{X}||\mathcal{Y}|}{4n\eta^2} \ .$$

PROOF. See Csiszár and Körner [17]. △

LEMMA 4.2. *For $\hat{P}_X \in \mathcal{P}_n(\mathcal{X})$, $x^n \in \hat{T}_X$, and $\hat{P}_Y \in \mathcal{P}_n(\mathcal{Y})$,*

$$\hat{T}_Y = \bigcup_{\substack{\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}): \\ \breve{P}_X = \hat{P}_X,\ \breve{P}_Y = \hat{P}_Y}} \breve{T}_{Y|X}(x^n) \ .$$

PROOF. Pick $y^n \in \hat{T}_Y$. Then $(x^n, y^n) \in \breve{T}_{XY}$ for some $\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ with $\breve{P}_X = \hat{P}_X$ and $\breve{P}_Y = \hat{P}_Y$. This implies that $y^n \in \breve{T}_{Y|X}(x^n)$. On the other hand, if $y^n \in \breve{T}_{Y|X}(x^n)$ for some $\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ with $\breve{P}_X = P_X$ and $\breve{P}_Y = P_Y$, then $y^n \in \breve{T}_Y = \hat{T}_Y$. △

LEMMA 4.3. *If $x^n \in \hat{T}_X$ and P_Y is any distribution, then*

$$T_{Y,\eta} = \bigcup_{\substack{\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}): \\ \breve{P}_X = \hat{P}_X,\ ||\breve{P}_Y - P_Y|| \le \eta,\ \breve{P}_Y \ll P_Y}} \breve{T}_{Y|X}(x^n) \ .$$

PROOF. From Lemma 4.2, we have

$$\hat{T}_Y = \bigcup_{\substack{\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y}): \\ \breve{P}_X = \hat{P}_X,\ \breve{P}_Y = \hat{P}_Y}} \breve{T}_{Y|X}(x^n) \ .$$

Since

$$T_{Y,\eta} = \bigcup_{||\hat{P}_Y - P_Y|| \le \eta,\ \hat{P}_Y \ll P_Y} \hat{T}_Y \ ,$$

we obtain

$$T_{Y,\eta} = \bigcup_{\substack{\breve{P}_{XY} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y}): \\ \breve{P}_X = \hat{P}_X,\ ||\breve{P}_Y - P_Y|| \le \eta,\ \breve{P}_Y \ll P_Y}} \breve{T}_{Y|X}(x^n) \ . \qquad \triangle$$

In a similar way we can prove the following two lemmas:

LEMMA 4.4. *For $\hat{P}_{UX} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X})$, $(u^n, x^n) \in \hat{T}_{UX}$, and $\hat{P}_{Y|U} \in \mathcal{P}_n(\mathcal{Y}|u^n)$,*

$$\hat{T}_{Y|U}(u^n) = \bigcup_{\substack{\breve{P}_{UXY} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}): \\ \breve{P}_{UX} = \hat{P}_{UX},\ \breve{P}_{UY} = \hat{P}_{UY}}} \breve{T}_{Y|UX}(u^n, x^n) \ .$$

Furthermore, if $P_{Y|U}$ is a stochastic matrix on $\mathcal{U}\times\mathcal{Y}$, then

$$T_{Y|U,\eta}(u^n) = \bigcup_{\substack{\breve{P}_{UXY} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}): \breve{P}_{UX} = \hat{P}_{UX}, \\ ||\breve{P}_{UY} - P_{Y|U}\hat{P}_U|| \le \eta,\ \breve{P}_{Y|U} \ll P_{Y|U}}} \breve{T}_{Y|UX}(u^n, x^n) \ . \qquad \triangle$$

LEMMA 4.5. *For $u^n \in \hat{T}_U$, $\hat{P}_{X|U} \in \mathcal{P}_n(\mathcal{X}|u^n)$, and $\hat{P}_{Y|U} \in \mathcal{P}_n(\mathcal{Y}|u^n)$,*

$$\hat{T}_{X|U}(u^n) \times \hat{T}_{Y|U}(u^n) = \bigcup_{\substack{\breve{P}_{UXY} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}): \\ \breve{P}_{UX} = \hat{P}_{UX},\ \breve{P}_{UY} = \hat{P}_{UY}}} \breve{T}_{XY|U}(u^n) \ . \qquad \triangle$$

(d) The coding technique. The following lemma is the basis for the coding procedure used in the proof of Theorem 4.1.

LEMMA 4.6. *Let k be an integer, and U be a simple random variable which is jointly distributed with X^k and Y^k, and is such that the joint distribution $P_{UX^kY^k}$ satisfies the Markov condition $P_{Y^k|UX^k} = P_{Y^k|X^k}$. Given $\eta > 0$, $\lambda > 0$, and $\delta > 0$, then for sufficiently large integers l and $n = kl$ there exists an integer M_n, a sequence $\{u_i^l\}_{i=1}^{M_n}$, and disjoint sets $\{C_i\}_{i=1}^{M_n} \subset \mathcal{X}^n$ such that*

$$M_n \leq \exp[l(I(U \wedge X^k) + \delta)] \,, \tag{2.3}$$

$$\{u_i^l\}_{i=1}^{M_n} \subset T_{U,\eta}^l \,, \tag{2.4}$$

$$C_i \subset T_{X^k|U,\eta}^l(u_i^l) \,, \tag{2.5}$$

and

$$\sum_{i=1}^{M_n} P_{X^kY^k}^l(C_i \times T_{Y^k|U,\bar{\eta}}^l(u_i^l)) \geq 1 - \lambda \,, \tag{2.6}$$

where $\bar{\eta}$ is a fixed multiple of η. Here, $I(U \wedge X^k)$ denotes the mutual information between U and X^k under the distribution P_{UX^k}, i.e.,

$$I(U \wedge X^k) = D(P_{UX^k} || P_U \times P_{X^k}) \,.$$

PROOF. See Han [11], Lemma 4. △

4.3. The main results.

In Theorem 4.1 we derive a lower bound on the error exponent $\theta(R, \epsilon)$ in terms of auxiliary distributions $P_{UX^kY^k}$ and $\tilde{P}_{UX^kY^k}$, where

(i) $P_{UX^kY^k}$ satisfies the Markov condition $P_{Y^k|UX^k} = P_{Y^k|X^k}$ and the inequality constraint $\frac{1}{k} D(P_{UX^k} || P_U \times P_{X^k}) \leq R$; and

(ii) $\tilde{P}_{UX^kY^k}$ satisfies the marginal constraints $\tilde{P}_{UX^k} = P_{UX^k}$, $\tilde{P}_{UY^k} = P_{UY^k}$.

With slight abuse of notation, we rewrite (i) and (ii) above as

(i′) $U \in \mathcal{S}(R, X^k Y^k)$, where

$$\mathcal{S}(R, X^k Y^k) \stackrel{\text{def}}{=} \{U : U \to X^k \to Y^k,\ \frac{1}{k} I(U \wedge X^k) \leq R\} \ ; \tag{3.1}$$

and

(ii′) $\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)$, where

$$\mathcal{L}_k(U) \stackrel{\text{def}}{=} \{\tilde{P}_{UX^kY^k} : \tilde{P}_{UX^k} = P_{UX^k},\ \tilde{P}_{UY^k} = P_{UY^k}\} \ . \tag{3.2}$$

Finally, we define

$$\begin{aligned} \theta_k(R) \stackrel{\text{def}}{=}\ & D(P_X || Q_X) \\ & + \frac{1}{k} \sup_{U \in \mathcal{S}(R, X^k Y^k)} \min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} D(\tilde{P}_{Y^k|UX^k} || Q_{Y^k|X^k} | \tilde{P}_{UX^k}) \ . \end{aligned} \tag{3.3}$$

The statement of the theorem is as follows:

THEOREM 4.1. *If* $R > 0$, $\epsilon \in (0, 1)$, *and* $Q_{XY} > 0$, *then*

$$\theta(R, \epsilon) \geq \sup_k \theta_k(R) \ .$$

PROOF. Fix $k \geq 1$ and let $U \in \mathcal{S}(R, X^k Y^k)$. Take l be sufficiently large for the statement of Lemma 4.6 to hold, and let $n = kl$. Hence, there exist M_n, $\{u_i^l\}_{i=1}^{M_n} \subset T_{U,\eta}^l$, and disjoint sets $\{C_i\}_{i=1}^{M_n}$ satisfying equations (2.3)–(2.6). Define the acceptance region $\mathcal{A}_n \subset \mathcal{X}^n \times \mathcal{Y}^n$ as

$$\mathcal{A}_n = \bigcup_{i=1}^{M_n} C_i \times T_{Y^k|U,\bar{\eta}}^l(u_i^l) \ .$$

Thus our encoding procedure entails transmiting the index i, where $x^n \in C_i$. Lemma 4.6 asserts that the code rate and the probability of correctly accepting the null hypothesis satisfy

$$\frac{1}{n}\log M_n \leq \frac{1}{k}(I(U \wedge X^k) + \delta) \leq R + \frac{\delta}{k}$$

and

$$P_{XY}^n(\mathcal{A}_n) \geq 1 - \lambda \geq 1 - \epsilon \ ,$$

respectively. The last inequality holds because $\lambda > 0$ can be chosen arbitrarily. To estimate the type II error rate, we can write

$$\begin{aligned} Q_{XY}^n(\mathcal{A}_n) &= \sum_{i=1}^{M_n} Q_{X^kY^k}^l(C_i \times T_{Y^k|U,\bar{\eta}}^l(u_i^l)) \\ &\leq \sum_{i=1}^{M_n} Q_{X^kY^k}^l(T_{X^k|U,\eta}^l(u_i^l) \times T_{Y^k|U,\bar{\eta}}^l(u_i^l)) \ . \end{aligned}$$

It follows from Lemma 4.5 that

$$Q_{XY}^n(\mathcal{A}_n) \leq \sum_{i=1}^{M_n} Q_{X^kY^k}^l\Big(\bigcup_{\hat{P}_{UX^kY^k} \in \Phi} \hat{T}_{X^kY^k|U}^l(u_i^l)\Big) \ ,$$

where, for ζ being a suitable multiple of η,

$$\Phi \stackrel{\text{def}}{=} \{\hat{P}_{UX^kY^k} \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X}^k \times \mathcal{Y}^k): \\ ||\hat{P}_{UX^k} - P_{UX^k}|| \leq \zeta, \ ||\hat{P}_{UY^k} - P_{UY^k}|| \leq \zeta\} \ .$$

Next we note that if $Q_{UX^kY^k} \stackrel{\text{def}}{=} P_U \times Q_{X^kY^k}$, then

$$Q_{X^kY^k}^l(\hat{T}_{X^kY^k|U}^l(u_i^l)) = Q_{X^kY^k|U}^l(\hat{T}_{X^kY^k|U}^l(u_i^l)|u_i^l)$$

and by Lemma 4.1,

$$Q_{X^kY^k}^l(\hat{T}_{X^kY^k|U}^l(u_i^l)) \leq \exp[-lD(\hat{P}_{X^kY^k|U}||Q_{X^kY^k}|\hat{P}_U)] \ .$$

As usual, it follows from continuity of the divergence functional that

$$Q_{XY}^n(\mathcal{A}_n) \leq \sum_{i=1}^{M_n} \exp[-l \min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} (D(\tilde{P}_{X^kY^k|U}||Q_{X^kY^k}|\tilde{P}_U) - \nu(\eta))] ,$$

where $\nu(\eta) \to 0$ as $\eta \to 0$. Hence

$$\begin{aligned} -\frac{1}{n}\log & Q_{XY}^n(\mathcal{A}_n) \\ & \geq \frac{1}{k}\{\min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} D(\tilde{P}_{X^kY^k|U}||Q_{X^kY^k}|\tilde{P}_U) - I(U \wedge X^k) - \nu(\eta) - \delta\} \\ & = D(P_X||Q_X) \\ & \quad + \frac{1}{k} \min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} D(\tilde{P}_{Y^k|UX^k}||Q_{Y^k|X^k}|\tilde{P}_{UX^k}) - \frac{\nu(\eta)}{k} - \frac{\delta}{k} . \end{aligned}$$

The above inequality is true for n integer multiple of k, and can be modified for arbitrary n by using a multiplicative factor equal to $\lfloor n/k \rfloor / \lceil n/k \rceil$. For every k, this factor clearly tends to unity as $n \to \infty$, and hence

$$\begin{aligned} \theta(R,\epsilon) & \geq \liminf_n (-\frac{1}{n}\log \beta_n(R,\epsilon)) \\ & \geq \liminf_n (-\frac{1}{n}\log Q_{XY}^n(\mathcal{A}_n)) \\ & \geq D(P_X||Q_X) + \frac{1}{k} \min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} D(\tilde{P}_{Y^k|UX^k}||Q_{Y^k|X^k}|\tilde{P}_{UX^k}) . \end{aligned}$$

Since the above is true for any $U \in \mathcal{S}(R, X^kY^k)$, we obtain $\theta(R,\epsilon) \geq \theta_k(R)$. $\triangle$

To show that the above lower bound is tight, we prove the following converse result.

THEOREM 4.2. *If $R > 0$, $\epsilon \in (0,1)$, and $Q_{XY} > 0$, then*

$$\theta(R,\epsilon) \leq \liminf_k \theta_k(R) .$$

PROOF. $\theta_k(R)$ defined in (3.3) can be lower bounded using the log-sum inequality as follows:

$$\begin{aligned}
\theta_k(R) &= D(P_X||Q_X) \\
&\quad + \frac{1}{k} \sup_{U \in \mathcal{S}(R, X^k Y^k)} \min_{\tilde{P}_{UX^kY^k} \in \mathcal{L}_k(U)} D(\tilde{P}_{Y^k|UX^k}||Q_{Y^k|X^k}|\tilde{P}_{UX^k}) \\
&= \frac{1}{k} \sup_{U \in \mathcal{S}(R, X^k Y^k)} \min_{\substack{\tilde{P}_{UX^kY^k}: \\ \tilde{P}_{UX^k}=P_{UX^k}, \\ \tilde{P}_{UY^k}=P_{UY^k}}} \sum_{u, x^k, y^k} \tilde{P}_{UX^kY^k}(u, x^k, y^k) \\
&\qquad\qquad \times \log \frac{\tilde{P}_{UX^kY^k}(u, x^k, y^k)}{Q_{X^kY^k}(x^k, y^k) P_{U|X^k}(u|x^k)} \\
&\geq \frac{1}{k} \sup_{U \in \mathcal{S}(R, X^k Y^k)} \sum_{u, y^k} P_{UY^k}(u, y^k) \log \frac{P_{UY^k}(u, y^k)}{\sum_{x^k} Q_{X^kY^k}(x^k, y^k) P_{U|X^k}(u|x^k)} .
\end{aligned}$$

If $U = f(X^k)$, where f is any function such that $H(f(X^k)) \leq kR$, then $f(X^k) \to X^k \to Y^k$, $I(f(X^k) \wedge X^k) = H(f(X^k))$, and hence $U \in \mathcal{S}(R, X^k Y^k)$. Therefore

$$\begin{aligned}
\theta_k(R) \;\geq\; \frac{1}{k} \sup_{f: H(f(X^k)) \leq kR} \sum_{y^k, u \in f(\mathcal{X}^k)} & P_{f(X^k)Y^k}(u, y^k) \times \\
& \log \frac{P_{f(X^k)Y^k}(u, y^k)}{\sum_{x^k} Q_{X^kY^k}(x^k, y^k) P_{f(X^k)|X^k}(u|x^k)} .
\end{aligned}$$

Noting that $P_{f(X^k)|X^k}(u|x^k) = I_{[f^{-1}\{u\}]}(x^k)$, where $I_A(\cdot)$ denotes the indicator function, we obtain

$$\begin{aligned}
&\theta_k(R) \\
&\quad \geq \frac{1}{k} \sup_{f: H(f(X^k)) \leq kR} \sum_{y^k, u \in f(\mathcal{X}^k)} P_{f(X^k)Y^k}(u, y^k) \log \frac{P_{f(X^k)Y^k}(u, y^k)}{\sum_{x^k \in f^{-1}\{u\}} Q_{X^kY^k}(x^k, y^k)} \\
&\quad = \frac{1}{k} \sup_{f: H(f(X^k)) \leq kR} \sum_{y^k, u \in f(\mathcal{X}^k)} P_{f(X^k)Y^k}(u, y^k) \log \frac{P_{f(X^k)Y^k}(u, y^k)}{Q_{f(X^k)Y^k}(u, y^k)} \\
&\quad = \frac{1}{k} \sup_{f: H(f(X^k)) \leq kR} D(P_{f(X^k)Y^k}||Q_{f(X^k)Y^k}) \\
&\quad \geq \frac{1}{k} \sup_{f: \log|f(\mathcal{X}^k)| \leq kR} D(P_{f(X^k)Y^k}||Q_{f(X^k)Y^k}) . \qquad (3.4)
\end{aligned}$$

It was shown in [10], Theorem 6 that the error exponent $\theta(R,\epsilon)$ is given by

$$\theta(R,\epsilon) = \lim_{k\to\infty} \frac{1}{k} \sup_{f:\log|f(\mathcal{X}^k)|\le kR} D(P_{f(X^k)Y^k}||Q_{f(X^k)Y^k}) \ .$$

We conclude that $\liminf_k \theta_k(R) \ge \theta(R,\epsilon)$. $\triangle$

THEOREM 4.3. *If $R>0$, $\epsilon\in(0,1)$, and $Q_{XY}>0$, then*

$$\theta(R,\epsilon) \ = \ \sup_k \theta_k(R) \ = \ \lim_k \theta_k(R) \ . \qquad \triangle$$

Theorem 4.4 asserts that if the alternative distribution is a product measure, then the error exponent given in Theorem 4.3 above is completely characterized by $\theta_1(R)$, i.e., $\theta(R,\epsilon)=\theta_1(R)$. Example 4.1, however, demonstrates that this is not true in general. Since $\theta_1(R)$ is also equal to the lower bound $\theta_L(R)$ derived by Han [11], Example 4.1 also shows that the said bound is not tight.

THEOREM 4.4. *If $R>0$, $\epsilon\in(0,1)$, and $Q_{XY}=Q_X\times Q_Y$ with $Q_X>0$ and $Q_Y>0$, then*

$$\theta(R,\epsilon)=\theta_1(R)=D(P_X||Q_X)+D(P_Y||Q_Y)+\max_{\substack{U:U\to X\to Y\\ I(U\wedge X)\le R,\ |\mathcal{U}|\le|\mathcal{X}|+1}} I(U\wedge Y) \ .$$

PROOF. We generalize the result obtained in [10] for $Q_X=P_X$ and $Q_Y=P_Y$, i.e., for testing a bivariate hypothesis against independence of marginals.

Setting $Q_{Y^k|X^k}=Q_Y^k$ in equation (3.3), we obtain

$$\theta_k(R)=D(P_X||Q_X) + \frac{1}{k}\sup_{\substack{U:U-X^k-Y^k,\\ I(U\wedge X^k)\le kR}} \min_{\substack{\tilde P_{UX^kY^k}:\\ \tilde P_{UX^k}=P_{UX^k},\ \tilde P_{UY^k}=P_{UY^k}}} D(\tilde P_{Y^k|UX^k}||Q_{Y^k}|\tilde P_{UX^k}) \ .$$

It is easy to check that $\tilde{P}_{UX^kY^k} = P_U \times P_{X^k|U} \times P_{Y^k|U}$ achieves the above minimum. Hence

$$\begin{aligned} D(\tilde{P}_{Y^k|UX^k}||Q_{Y^k}|\tilde{P}_{UX^k}) &= \sum_{u,x^k,y^k} \tilde{P}_{UX^kY^k}(u,x^k,y^k) \log \frac{P_{Y^k|U}(y^k|u)}{Q_{Y^k}(y^k)} \\ &= \sum_{u,y^k} P_{UY^k}(u,y^k) \log \frac{P_{Y^k}(y^k)P_{Y^k|U}(y^k|u)}{P_{Y^k}(y^k)Q_{Y^k}(y^k)} \\ &= D(P_{Y^k}||Q_{Y^k}) + I(U \wedge Y^k) \ . \end{aligned}$$

$\theta(R,\epsilon)$ is thus equal to

$$D(P_X||Q_X) + D(P_Y||Q_Y) + \sup_k \sup_{\substack{U:U\to X^k\to Y^k, \\ I(U\wedge X^k)\le kR}} \frac{1}{k} I(U \wedge Y^k) \ .$$

Making use of the results of the entropy characterization problem [17, Chapter 3, Theorem 3.20], we obtain

$$\theta(R,\epsilon) = D(P_X||Q_X) + D(P_Y||Q_Y) + \sup_{\substack{U:U\to X\to Y, \\ I(U\wedge X)\le R}} I(U \wedge Y) = \theta_1(R) \ .$$

Finally, the range constraint on $|\mathcal{U}|$ in $\theta_1(R)$ can be derived by means of a standard convexity argument [17, Chapter 3, Lemmas 3.4 and 3.5]. $\triangle$

EXAMPLE 4.1. Let the null and alternative distributions be given by

$$P_{XY}: \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.5-p & p \\ p & 0.5-p \end{pmatrix} \end{array}, \qquad Q_{XY}: \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 3/8 & 1/8 \\ 1/8 & 3/8 \end{pmatrix} \end{array},$$

where $p \le 1/2$. We take $R = 0.5$ and note that $H(X) = 1$.

Our computation of $\theta_1(R)$ for selected values of p yields

$$\theta_1(R) = \begin{cases} 0.208, & \text{if } p = 0 \ ; \\ 0.206, & \text{if } p = 0.0001 \ ; \\ 0.199, & \text{if } p = 0.001 \ ; \\ 0.153, & \text{if } p = 0.01 \ . \end{cases}$$

Next we find lower bounds on $\theta_2(R)$ for the same selected values of p. We can write

$$P_{X^2Y^2} = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cccc} 00 & 01 & 10 & 11 \end{array} \\ \begin{pmatrix} (0.5-p)^2 & p(0.5-p) & p(0.5-p) & p^2 \\ p(0.5-p) & (0.5-p)^2 & p^2 & p(0.5-p) \\ p(0.5-p) & p^2 & (0.5-p)^2 & p(0.5-p) \\ p^2 & p(0.5-p) & p(0.5-p) & (0.5-p)^2 \end{pmatrix} \end{array} .$$

Let $U = f(X^2) = X_1 \oplus X_2$ (mod-2 sum of X_1 and X_2). Then $P_U(0) = P_{X^2}(00) + P_{X^2}(11) = 0.5$, which implies that $H(U) = 1$ and $I(U \wedge X^2) = H(U) - H(U|X^2) = H(U) = 1 = 2R$. Thus $U \in \mathcal{S}(0.5, X^2Y^2)$. We also have

$$P_{UY^2} = \begin{pmatrix} p^2 + (0.5-p)^2 & 2p(0.5-p) & 2p(0.5-p) & p^2 + (0.5-p)^2 \\ 2p(0.5-p) & p^2 + (0.5-p)^2 & p^2 + (0.5-p)^2 & 2p(0.5-p) \end{pmatrix} ,$$

and similarly,

$$Q_{UY^2} = \begin{pmatrix} 5/32 & 3/32 & 3/32 & 5/32 \\ 3/32 & 5/32 & 5/32 & 3/32 \end{pmatrix} .$$

It follows from equation (3.4) that

$$\theta_2(R) \geq \frac{1}{2} D(P_{UY^2} || Q_{UY^2}) = \begin{cases} 0.339, & \text{if } p = 0 \ ; \\ 0.337, & \text{if } p = 0.0001 \ ; \\ 0.322, & \text{if } p = 0.001 \ ; \\ 0.234, & \text{if } p = 0.01 \ . \end{cases}$$

This shows that $\theta(0.5, \epsilon) \geq \theta_2(0.5) > \theta_1(0.5)$ for the above selected values of p.

4.4. Concluding remarks.

Example 4.1 indicates that our characterization of the error exponent obtained in terms of a sequence of single-letter formulas does not in general reduce to one single-letterized formula, i.e., $\theta(R, \epsilon) \neq \theta_1(R)$. There are, however, special cases (cf. Theorem 4.4) where the error exponent is indeed given by $\theta_1(R)$.

The problem of finding a single-letter formula of $\theta(R, \epsilon)$ was termed a *divergence characterization problem* in [10]. Our results demonstrate that this problem is a special case of the single-letter characterization of the two-dimensional convex closure of $\bigcup_{k=1}^{\infty} \mathcal{H}_k$, where

$$\mathcal{H}_k \stackrel{\text{def}}{=} \Big\{ \big(\frac{1}{k} H(X^k|U), \min_{\substack{\tilde{P}_{UX^kY^k}: \\ \tilde{P}_{UX^k}=P_{UX^k}, \\ \tilde{P}_{UY^k}=P_{UY^k}}} \frac{1}{k} D(\tilde{P}_{Y^k|UX^k} || Q_{Y^k|X^k} | \tilde{P}_{UX^k})\big): \\ U \to X^k \to Y^k \Big\} .$$

This problem remains open to date.

CHAPTER 5

DISTRIBUTED DETECTION WITH FEEDBACK

5.1. Introduction.

In the previous chapters we studied the asymptotics of distributed detection systems with fixed numbers of sensors and growing sample sizes. In this chapter we consider the case where the size of the sample obtained by each sensor is held fixed, and the number of sensors approaches infinity.

Systems with a large number of sensors have been studied in [23,28,33]. Our work is thematically related to [23], in which a Bayesian multiple hypothesis testing problem with N sensors making independent observations was considered. Both sensors and fusion center used deterministic rules, and the optimization aimed at minimizing the overall probability of error. For binary hypothesis testing, under the assumption of identical statistics across sensors, it was shown that the sensors could use the same decision rule without loss of asymptotic optimality. Similar results were obtained in [28] for sequential data fusion.

We consider in this chapter two variations on this setup. One entails the transmission of sensor data to the fusion center in two stages, with broadcast of feedback information from the center to the sensors after the first stage. The other variation involves information exchange between sensors prior to transmission to the fusion center; this exchange is effected through a *feedback center*, which processes binary data from the sensors and thereafter broadcasts a single feedback bit back to the sensors. The latter variation is of relevance to situations where the channels between sensors and fusion center are of restricted capacity

and transmission across those channels entails considerable cost. We show that under the Neyman-Pearson criterion, only the latter type of feedback yields an improvement on the asymptotic performance of the system (as $N \to \infty$), and we derive the associated error exponents.

5.2. Problem definition and preliminaries.

We have two hypotheses H_0 and H_1, and N sensors $\{S_i\}_{i=1}^N$. Each sensor S_i observes a random variable X_i that takes values in a finite set $\mathcal{X}$. We assume that the sequence $X^N = (X_1, \ldots, X_N)$ is identically distributed and independent under both hypotheses. Thus if P_X and Q_X are the distributions of X_i under H_0 and H_1, respectively, we have for all $x^n \in \mathcal{X}^N$

$$P_{X^N}(x^N) = P_X^N(x^N) = \prod_{i=1}^N P_{X_i}(x_i) \ ,$$

$$Q_{X^N}(x^N) = Q_X^N(x^N) = \prod_{i=1}^N Q_{X_i}(x_i) \ .$$

We study four different systems for processing the data collected by the sensors:

System 1. No feedback is employed in this system. Each sensor transmits a random message $U_i \in \mathcal{U}$, where $\mathcal{U}$ is a finite set of cardinality at most $|\mathcal{X}|$. U_i is generated by means of a behavioral rule [34] or simply a random encoder, which can be represented by a conditional distribution $\Delta_{U_i|X_i}$ on $\mathcal{U} \times \mathcal{X}$. As soon as the fusion center C collects the local messages of each sensor $(U_1, \ldots, U_N)$, it declares hypothesis H_0 to be true if U^N lies in some acceptance region $\mathcal{A}_N \subset \mathcal{U}^N$. The optimal encoders $\{\Delta_{U_i|X_i}\}$ and acceptance region minimize the type II error

$Q_{U^N}(\mathcal{A}_N)$ subject to the constraint

$$P_{U^N}(\mathcal{A}_N^c) \leq \epsilon$$

on the type I error, where $\epsilon \in (0,1)$. The joint distribution P_{U^N} of the messages is given by

$$P_{U^N}(u^N) = \prod_{i=1}^{N} P_{U_i}(u_i) \ ,$$

where

$$P_{U_i}(u_i) = \sum_x \Delta_{U_i|X_i}(u_i|x) P_X(x) \ .$$

We will also represent the above by the vector-matrix product $P_{U_i} = P_X \Delta_{U_i|X_i}$, where P_{U_i} is a $|\mathcal{U}|$-dimensional row vector, P_X is a $|\mathcal{X}|$-dimensional row vector, and $\Delta_{U_i|X_i}$ is a $|\mathcal{U}| \times |\mathcal{X}|$ matrix. $Q_{U^N}(u^N)$ is defined in a similar manner. We denote the minimum probability of type II error by $\beta_N^{(1)}(|\mathcal{U}|, \epsilon)$, i.e.,

$$\beta_N^{(1)}(|\mathcal{U}|, \epsilon) \stackrel{\text{def}}{=} \inf_{\{\Delta_{U_i|X_i}\},\ \mathcal{A}_N \subset \mathcal{U}^N} \{Q_{U^N}(\mathcal{A}_N) \colon P_{U^N}(\mathcal{A}_N) \geq 1 - \epsilon\} \ .$$

We are interested in the asymptotic behavior of $\beta_N^{(1)}(|\mathcal{U}|, \epsilon)$ as $N \to \infty$. The resulting error exponent is given by

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) \stackrel{\text{def}}{=} -\lim_N \frac{1}{N} \log \beta_N^{(1)}(|\mathcal{U}|, \epsilon) \ ,$$

provided the limit exists.

In investigating the effect of feedback on the above system, we study two patterns of information flow within the system:

(i) The first pattern entails two-stage transmission of a two-bit message using a one-bit feedback packet from the fusion center.

(ii) The second pattern involves partial information transmission to the feedback center which then provides a one-bit message to all sensors. Based on this

feedback information, the sensors re-encode their observations into one-bit messages and communicate them to the fusion center.

System 2 described below follows the first information flow pattern, whereas Systems 3 and 4 follow the second pattern.

System 2. Here we assume that the ith sensor transmits binary messages in two stages: message U_i in the first stage, followed by message W_i in the second. After the first stage, and based on the received sequence $U_1, \ldots, U_N$, the fusion center produces a binary feedback message V, and communicates it to the sensors. This bit is 0 if U^N lies in some region $\mathcal{C}_N \subset \mathcal{U}^N$ called the *feedback acceptance region*, and 1 otherwise. Each sensor S_i then generates two binary messages Y_i and Z_i (that take values in the same binary alphabet $\mathcal{W}$) according to distributions $\Delta_{Y_i|U_iX_i}$ and $\Delta_{Z_i|U_iX_i}$, respectively, and uses the feedback bit to decide which of Y_i, Z_i to transmit to the fusion center as the second information bit W_i. Thus

$$W_i \stackrel{\text{def}}{=} Y_i I_{[V=0]} + Z_i I_{[V=1]} \,, \tag{2.1}$$

where I_A denotes the indicator function of the event A. Since V depends on the observations of all sensors, the W_i's are in general dependent. The fusion center C collects these messages W^N and uses these along with U^N to declare that H_0 is true if $U^N W^N$ lies in an acceptance region $\mathcal{A}_N$ which is a subset of $\mathcal{U}^N \times \mathcal{W}^N$. The optimal encoders $\{\Delta_{U_iY_i|X_i},\ \Delta_{U_iZ_i|X_i}\}$, feedback acceptance region $\mathcal{C}_N$, and acceptance region $\mathcal{A}_N$ are those that minimize the type II error, $Q_{U^N W^N}(\mathcal{A}_N)$, subject to the constraint

$$P_{U^N W^N}(\mathcal{A}_N^c) \leq \epsilon$$

on the type I error. By (2.1), for all $(u^N, w^N) \in \mathcal{U}^N \times \mathcal{W}^N$, we can write

$$P_{U^N W^N}(u^N, w^N) = \begin{cases} P_{U^N Y^N}(u^N, w^N), & \text{if } u^N \in \mathcal{C}_N; \\ P_{U^N Z^N}(u^N, w^N), & \text{otherwise,} \end{cases} \tag{2.2}$$

where

$$P_{U^N Y^N}(u^N, w^N) = \prod_{i=1}^{N} P_{U_i Y_i}(u_i, w_i), \qquad P_{U_i Y_i} = P_X \Delta_{U_i Y_i | X_i} \ ,$$

and

$$P_{U^N Z^N}(u^N, w^N) = \prod_{i=1}^{N} P_{U_i Z_i}(u_i, w_i), \qquad P_{U_i Z_i} = P_X \Delta_{U_i Z_i | X_i} \ .$$

Hence, by (2.2)

$$P_{U^N W^N}(\mathcal{A}_N) = P_{U^N Y^N}(\mathcal{A}_N \cap (\mathcal{C}_N \times \mathcal{W}^N)) + P_{U^N Z^N}(\mathcal{A}_N \cap (\mathcal{C}_N^c \times \mathcal{W}^N)) \ .$$

$Q_{U^N W^N}$ is evaluated in a similar way. The optimal type II error is defined as follows

$$\beta_N^{(2)}(\epsilon) \stackrel{\text{def}}{=} \inf_{\substack{\{\Delta_{U_i Y_i | X_i},\ \Delta_{U_i Z_i | X_i}\}, \\ \mathcal{C}_N \subset \mathcal{U}^N,\ \mathcal{A}_N \subset \mathcal{U}^N \times \mathcal{W}^N}} \{Q_{U^N W^N}(\mathcal{A}_N) : P_{U^N W^N}(\mathcal{A}_N) \geq 1 - \epsilon\} \ .$$

The corresponding error exponent is given by

$$\theta^{(2)}(\epsilon) \stackrel{\text{def}}{=} -\lim_N \frac{1}{N} \log \beta_N^{(2)}(\epsilon) \ .$$

System 3. This system differs from System 2 in that here we have two distinct centers: the feedback center $\bar{C}$ and the fusion center C. The sensors transmit the first messages U^N to $\bar{C}$, which broadcasts a binary feedback message V (generated in exactly the same manner as in System 2) to all sensors and to C. The sensors then transmit the second messages W^N (generated as in System

2) to C. C uses W^N along with V to declare the final decision. Therefore, the acceptance region $\mathcal{A}_N \subset \mathcal{U}^N \times \mathcal{W}^N$ can be written as the disjoint union

$$\mathcal{A}_N = (\mathcal{C}_N \times \mathcal{F}_N) \cup (\mathcal{C}_N^c \times \mathcal{E}_N) \ ,$$

where $\mathcal{C}_N \subset \mathcal{U}^N$ and $\mathcal{F}_N, \mathcal{E}_N \subset \mathcal{W}^N$. With the aid of (2.2), we can write

$$P_{U^N W^N}(\mathcal{A}_N) = P_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) + P_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \ .$$

The corresponding optimal type II error is thus

$$\beta_N^{(3)}(\epsilon) \stackrel{\text{def}}{=} \inf_{\substack{\{\Delta_{U_i Y_i|X_i},\ \Delta_{U_i Z_i|X_i}\}, \\ \mathcal{C}_N \subset \mathcal{U}^N,\ \mathcal{F}_N,\ \mathcal{E}_N \subset \mathcal{W}^N}} \{Q_{U^N W^N}(\mathcal{A}_N): \\ \mathcal{A}_N = (\mathcal{C}_N \times \mathcal{F}_N) \cup (\mathcal{C}_N^c \times \mathcal{E}_N),\ P_{U^N W^N}(\mathcal{A}_N) \geq 1 - \epsilon\}$$

and the error exponent is given by

$$\theta^{(3)}(\epsilon) \stackrel{\text{def}}{=} -\lim_N \frac{1}{N} \log \beta_N^{(3)}(\epsilon) \ .$$

System 4. This system differs slightly from System 3 in that the two centers $\bar{C}$, C do not communicate. Thus C uses W^N only to determine the true hypothesis. The statement of the problem is summarized as follows. The acceptance region used by C is $\mathcal{A}_N$, which is a subset of $\mathcal{W}^N$, not $\mathcal{U}^N$ as in System 1. The type I and type II errors are given by

$$P_{W^N}(\mathcal{A}_N) = P_{U^N Y^N}(\mathcal{C}_N \times \mathcal{A}_N) + P_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{A}_N)$$

and

$$Q_{W^N}(\mathcal{A}_N) = Q_{U^N Y^N}(\mathcal{C}_N \times \mathcal{A}_N) + Q_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{A}_N) \ ,$$

respectively. The optimal type II error is thus

$$\beta_N^{(4)}(\epsilon) \stackrel{\text{def}}{=} \inf_{\substack{\{\Delta_{U_i Y_i|X_i},\ \Delta_{U_i Z_i|X_i}\},\\ \mathcal{C}_N \subset \mathcal{U}^N,\ \mathcal{A}_N \subset \mathcal{W}^N}} \{Q_{W^N}(\mathcal{A}_N) : P_{W^N}(\mathcal{A}_N) \geq 1 - \epsilon\}$$

and the error exponent is given by

$$\theta^{(4)}(\epsilon) \stackrel{\text{def}}{=} -\lim_N \frac{1}{N} \log \beta_N^{(4)}(\epsilon) \ .$$

To investigate the effect of the first kind of feedback, we compare the error exponent in System 1 (when $|\mathcal{U}| = 4$) with that in System 2. The effect of the second kind of feedback is illustrated by comparing the error exponents in System 1 (when $|\mathcal{U}| = 2$) with that in Systems 3 and 4.

It turns out that System 2 performs similarly to System 1 (with $|\mathcal{U}| = 4$). This implies that the first kind of feedback does not improve the error exponent in System 1, but it is certainly useful in reducing the complexity of the setup. On the other hand Systems 3 and 4 outperform System 1 (with $|\mathcal{U}| = 2$) in general, hence the second kind of feedback is useful. We will see also that the error exponent in System 3 is equal to that in System 4, i.e., C does not need to know the feedback bit transmitted by $\bar{C}$.

In Section 5.3 we evaluate the error exponents of the above systems under the assumption that all sensors use the same encoder, and show that the sensors can employ only deterministic encoders (i.e., encoders described by deterministic behavioral rules) without loss of asymptotic optimality. In Section 5.4 we evaluate the error exponents in Systems 1 and 2 assuming that the sensors are not restricted to use the same encoder.

(c) Typical sequences. The basic definitions and properties of typical sequences cited in Section 2.2 will be used here with N replacing n.

For convenience we restate here Lemmas 2.1 and 2.2 with slight modifications.

LEMMA 5.1. *For any* $\hat{P}_X$ *in* $\mathcal{P}_N(\mathcal{X})$, *and* Q_X *in* $\mathcal{P}(\mathcal{X})$

$$(N+1)^{-|\mathcal{X}|}\exp[NH(\hat{P}_X)] \leq |\hat{T}_X^N| \leq \exp[NH(\hat{P}_X)] ,$$

and

$$(N+1)^{-|\mathcal{X}|}\exp[-ND(\hat{P}_X||Q_X)] \leq Q_X^N(\hat{T}_X^N) \leq \exp[-ND(\hat{P}_X||Q_X)] .$$

LEMMA 5.2. *For any two distributions* P_X, Q_X *on* $\mathcal{X}$, *and* $\eta > 0$,

$$P_X^N(T_{X,\eta}^N) \geq 1 - \frac{|\mathcal{X}|}{4N\eta^2} ,$$

$$Q_X^N(T_{X,\eta}^N) \leq \exp[-N(D(P_X||Q_X) - \delta_N - \nu(\eta))] ,$$

where $\delta_N = \frac{|\mathcal{X}|\log(N+1)}{N} \to 0$, *and* $\nu(\eta) \to 0$ *as* $\eta \to 0$.

We will need the following lemma.

LEMMA 5.3. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be any binary sets. Fix* $\rho > 0$, $\delta \in (0,1)$. *Then there exists a sequence*

$$\nu_N = \nu_N(\rho, \delta, |\mathcal{X}|, |\mathcal{Y}|) \to 0$$

such that for every P_{XY}, $Q_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $C \in \mathcal{X}^N$, $F \in \mathcal{Y}^N$ *satisfying*

$$\min_{x,y:Q_{XY}(x,y)>0} Q_{XY}(x,y) > \rho ,$$

$$D(P_{XY}||Q_{XY}) < \infty ,$$

and

$$P_{XY}^N(C \times F) \geq \delta \ ,$$

the following is true:

$$Q_{XY}^N(C \times F) \geq \exp[-N(d(P_X, P_Y||Q_{XY}) + \nu_N)] \ ,$$

where

$$d(P_X, P_Y||Q_{XY}) \stackrel{\text{def}}{=} \min_{\substack{\tilde{P}_{XY} \\ \tilde{P}_X = P_X, \ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY}||Q_{XY}) \ .$$

PROOF. *Case 1.* $Q_{XY} > 0$. The statement follows immediately from Theorem 2.3 and is also true for arbitrary $\mathcal{X}, \mathcal{Y}$.

Case 2. Q_{XY} has zeros. The binary assumption on the sets $\mathcal{X}$ and $\mathcal{Y}$ is critical here. If the distribution $\tilde{P}_{XY}$ achieves the minimum in $d(P_X, P_Y||Q_{XY})$, then $\tilde{P}_{XY} \ll Q_{XY}$ (otherwise the divergence equals infinity). The constraints $\tilde{P}_X = P_X$, $\tilde{P}_Y = P_Y$ force $\tilde{P}_{XY}$ to be identical to P_{XY}, i.e., here $d(P_X, P_Y||Q_{XY}) = D(P_{XY}||Q_{XY})$. Stein's lemma [8] ensures the existence of a sequence $\lambda_N \to 0$ such that

$$Q_{XY}^N(C \times F) \ \geq \ \exp[-N(D(P_{XY}||Q_{XY}) + \lambda_N)] \ . \qquad \triangle$$

In the following sections we will omit the superscript N from T, as N will be essentially constant.

5.3. The main results.

The first theorem is a straightforward adaptation of Stein's lemma for which we have been unable to find a reference in the literature; hence we give a full proof.

THEOREM 5.1. *The error exponent for System 1, assuming all sensors use the same encoder and* $|\mathcal{U}| \leq |\mathcal{X}|$, *is given by*

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) = \sup_{\substack{\Delta_{U|X}: \\ P_U = P_X \Delta_{U|X},\ Q_U = Q_X \Delta_{U|X}}} D(P_U || Q_U)$$

for all $\epsilon \in (0, 1)$.

PROOF. *Direct part.* If $|\mathcal{U}| = |\mathcal{X}|$, then the error exponent is given by Stein's lemma. Let $\mathcal{U}$ be a set of cardinality not greater than $|\mathcal{X}|$. We assume all sensors use the same encoder, i.e., $\Delta_{U_i|X_i} = \Delta_{U|X}$, for all $1 \leq i \leq N$. We thus have

$$P_{U_i} = P_U = P_X \Delta_{U|X}, \qquad Q_{U_i} = Q_U = Q_X \Delta_{U|X} \ .$$

Set the acceptance region

$$\mathcal{A}_N = T_{U,\eta} \ ,$$

where $\eta > 0$ is arbitrary small. Then from Lemma 5.2,

$$P_{U^N}(\mathcal{A}_N) = P_U^N(\mathcal{A}_N) \geq 1 - \frac{|\mathcal{U}|}{4N\eta^2} \ ,$$

which is greater than $1 - \epsilon$ if N is large enough. The type II error is upper bounded by $\exp[-N(D(P_U||Q_U) - \delta_N - \nu(\eta))]$, where $\delta_N \to 0$ and $\nu(\eta) \to 0$ as $\eta \to 0$ (cf. Lemma 5.2). Since the conditional distribution $\Delta_{U|X}$ is arbitrary, we have

$$\beta_N^{(1)}(|\mathcal{U}|, \epsilon) \leq \inf_{\Delta_{U|X}} \exp[-N(D(P_U||Q_U) - \delta_N - \nu(\eta))] \ .$$

By definition of the error exponent,

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) \geq \sup_{\Delta_{U|X}} D(P_U||Q_U) - \nu(\eta) \ .$$

Since $\eta > 0$ is arbitrary, the proof of the direct part is complete.

Converse part. Assume that all sensors use the same encoder $\Delta_{U|X}$. Let $\mathcal{A}_N \subset \mathcal{U}^N$ be any acceptance region satisfying the constraint

$$P_{U^N}(\mathcal{A}_N) \geq 1 - \epsilon \, .$$

Hence from Stein's lemma, for N large enough there exists a sequence $\lambda_N \to 0$, depending only on $|\mathcal{U}|$ and ϵ, such that

$$\begin{aligned} Q_{U^N}(\mathcal{A}_N) &\geq \exp[-N(D(P_U||Q_U) + \lambda_N)] \\ &\geq \exp[-N(\sup_{\Delta_{U|X}} D(P_U||Q_U) + \lambda_N)] \, . \end{aligned}$$

Since $\Delta_{U|X}$ was arbitrary, and $\mathcal{A}_N$ was any set satisfying the constraint on type I error, we obtain

$$\beta_N^{(1)}(|\mathcal{U}|, \epsilon) \geq \exp[-N(\sup_{\Delta_{U|X}} D(P_U||Q_U) + \lambda_N)] \, .$$

Thus

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) \leq \sup_{\Delta_{U|X}} D(P_U||Q_U) \, . \qquad \Delta$$

THEOREM 5.2. *The error exponent for System 2, assuming all sensors use the same pair of encoders, is given by*

$$\theta^{(2)}(\epsilon) = \sup_{\substack{\Delta_{UY|X}: \\ P_{UY} = P_X \Delta_{UY|X}, \; Q_{UY} = Q_X \Delta_{UY|X}}} D(P_{UY}||Q_{UY})$$

for all $\epsilon \in (0, 1)$.

REMARK. One can see that if $|\mathcal{U}| = 4$ in Theorem 5.1, then the error exponents in System 1 and 2 are exactly the same, thus in asymptotic terms,

the feedback bit used by the fusion center does not convey essential information to the sensors. As a matter of fact, one can show that the same error exponent prevails for any *fixed* number of feedback bits.

PROOF. *Direct part.* It is obvious that this system will do at least as well as the one with no feedback, hence $\theta^{(2)}(\epsilon) \geq \theta^{(1)}(\epsilon)$.

Converse part. Assume that all sensors use a pair of encoders $\Delta_{UY|X}$, $\Delta_{UZ|X}$ Let $\mathcal{C}_N \subset \mathcal{U}^N$, $\mathcal{A}_N \subset \mathcal{U}^N \times \mathcal{W}^N$ be any regions satisfying the constraint

$$\begin{aligned} P_{U^N W^N}(\mathcal{A}_N) &= P_{U^N Y^N}(\mathcal{A}_N \cap (\mathcal{C}_N \times \mathcal{W}^N)) + P_{U^N Z^N}(\mathcal{A}_N \cap (\mathcal{C}_N^c \times \mathcal{W}^N)) \\ &\geq 1 - \epsilon \, . \end{aligned}$$

Hence, either $P_{U^N Y^N}(\mathcal{A}_N \cap (\mathcal{C}_N \times \mathcal{W}^N)) \geq (1-\epsilon)/2$ or $P_{U^N Z^N}(\mathcal{A}_N \cap (\mathcal{C}_N^c \times \mathcal{W}^N)) \geq (1-\epsilon)/2$. Using the same method as in the proof of the converse part of Theorem 5.1, we obtain

$$\begin{aligned} \theta^{(2)}(\epsilon) &\leq \{ \sup_{\Delta_{UY|X}} D(P_{UY}||Q_{UY}) \} \vee \{ \sup_{\Delta_{UZ|X}} D(P_{UZ}||Q_{UZ}) \} \\ &= \sup_{\Delta_{UY|X}} D(P_{UY}||Q_{UY}) \, . \end{aligned}$$ △

THEOREM 5.3. *The error exponents for Systems 3 and 4, assuming all sensors use the same pair of encoders and* $D(P_X||Q_X) < \infty$, *are given by*

$$\theta^{(3)}(\epsilon) = \theta^{(4)}(\epsilon) = \sup_{\substack{\Delta_{UY|X}: \\ P_{UY} = P_X \Delta_{UY|X},\ Q_{UY} = Q_X \Delta_{UY|X}}} d(P_U, P_Y||Q_{UY})$$

for all $\epsilon \in (0,1)$.

PROOF. *Direct part.* By the problem statement, we have $\theta^{(3)}(\epsilon) \geq \theta^{(4)}(\epsilon)$. Hence it is enough to show the direct part for System 4 only. Pick an arbitrary

pair of conditional distributions $\Delta_{UY|X}$, $\Delta_{UZ|X}$ as fixed encoders for all sensors. Set $\mathcal{C}_N = T_{U,\eta}$, $\mathcal{A}_N = T_{Y,\eta}$, where $\eta > 0$ is arbitrary. For the type I error in System 4, we can write

$$\begin{aligned} P_{W^N}(\mathcal{A}_N^c) &= P_{U^N Y^N}(\mathcal{C}_N \times \mathcal{A}_N^c) + P_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{A}_N^c) \\ &\leq P_{Y^N}(\mathcal{A}_N^c) + P_{U^N}(\mathcal{C}_N^c) \\ &\leq \frac{|\mathcal{Y}|}{4N\eta^2} + \frac{|\mathcal{U}|}{4N\eta^2} \leq \epsilon \end{aligned}$$

if N is large enough. The Type II error in System 4 is upper bounded as follows:

$$\begin{aligned} Q_{W^N}(\mathcal{A}_N) &= Q_{U^N Y^N}(\mathcal{C}_N \times \mathcal{A}_N) + Q_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{A}_N) \\ &\leq Q_{U^N Y^N}(T_{U,\eta} \times T_{Y,\eta}) + Q_{Z^N}(T_{Y,\eta}) \\ &\leq \exp[-N(d(P_U, P_Y||Q_{UY}) - \delta_N - \nu(\eta))] + Q_{Z^N}(T_{Y,\eta}) \ . \end{aligned}$$

By a suitable choice of $\Delta_{Z|X}$ the last term in the above inequality can be made equal to zero. Indeed, for any fixed y_0 with $P_Y(y_0) > 0$ we can always choose a trivial encoder $\Delta_{Z|X}$ such that $\Delta_{Z|X}(y_0|x) = 0$ for all $x \in \mathcal{X}$. It follows that $Q_Z(y_0) = 0$. Since $P_Y(y_0) > 0$, there exists an $1 \geq i(y^N) \geq N$ such that $y_{i(y^N)} = y_0$ for each sequence y^N in $T_{Y,\eta}$. Thus $Q_{Z^N}(T_{Y,\eta}) = 0$. This, together with the fact that the encoders were arbitrarily chosen, yields

$$\theta^{(4)}(\epsilon) \geq \sup_{\Delta_{UY|X}} d(P_U, P_Y||Q_{UY}) - \nu(\eta) \ .$$

Converse part. It is enough to show the converse part for System 3 only, since $\theta^{(3)}(\epsilon) \geq \theta^{(4)}(\epsilon)$. Assume that all sensors use a pair of encoders $\Delta_{UY|X}, \Delta_{UZ|X}$. Note that $D(P_{UY}||Q_{UY}) < \infty$ and $D(P_{UZ}||Q_{UZ}) < \infty$ since $D(P_X||Q_X) < \infty$. Let $\mathcal{C}_N \subset \mathcal{U}^N$, $\mathcal{F}_N, \mathcal{E}_N \subset \mathcal{W}^N$, $\mathcal{A}_N = \mathcal{C}_N \times \mathcal{F}_N \cup \mathcal{C}_N^c \times \mathcal{E}_N$ be satisfying the type

I error constraint

$$P_{U^N W^N}(\mathcal{A}_N) \;=\; P_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) + P_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \;\geq 1-\epsilon\;.$$

Hence, either $P_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) \geq (1-\epsilon)/2$ or $P_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \geq (1-\epsilon)/2$. Using Lemma 5.3, we have either

$$\begin{aligned} Q_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) \;&\geq\; \exp[-N(d(P_U, P_Y||Q_{UY}) + \nu_N)] \\ &\geq\; \exp[-N(\sup_{\Delta_{UY|X}} d(P_U, P_Y||Q_{UY}) + \nu_N)]\;, \end{aligned}$$

or

$$\begin{aligned} Q_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \;&\geq\; \exp[-N(d(P_U, P_Z||Q_{UZ}) + \nu_N)] \\ &\geq\; \exp[-N(\sup_{\Delta_{UY|X}} d(P_U, P_Y||Q_{UY}) + \nu_N)]\;. \end{aligned}$$

This yields

$$\begin{aligned} Q_{U^N W^N}(\mathcal{A}_N) \;&=\; Q_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) + Q_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \\ &\geq\; Q_{U^N Y^N}(\mathcal{C}_N \times \mathcal{F}_N) \vee Q_{U^N Z^N}(\mathcal{C}_N^c \times \mathcal{E}_N) \\ &\geq\; \exp[-N(\sup_{\Delta_{UY|X}} d(P_U, P_Y||Q_{UY}) + \lambda_N)]\;, \end{aligned}$$

and therefore also

$$\theta^{(3)}(\epsilon) \leq \sup_{P_{UY|X}} d(P_U, P_Y||Q_{UY})\;. \qquad \triangle$$

REMARK. In general, the error exponent in Systems 3 and 4 is better than that in System 1 (with $|\mathcal{U}| = 2$). This is because $d(P_U, P_Y||Q_{UY}) \geq D(P_U||Q_U)$.

In what follows we will see that it is sufficient for all sensors to employ deterministic encoders in order to achieve the above error exponents. We first need to define $\Phi(\Delta_{U|X})$ and $\Psi(\Delta_{UY|X})$ as follows.

$$\Phi(\Delta_{U|X}) \stackrel{\text{def}}{=} D(P_U||Q_U)\;, \tag{3.1}$$

where

$$P_U = P_X \Delta_{U|X} \qquad \text{and} \qquad Q_U = Q_X \Delta_{U|X} \ , \tag{3.2}$$

and

$$\Psi(\Delta_{UY|X}) \stackrel{\text{def}}{=} d(P_U, P_Y || Q_{UY}) \ , \tag{3.3}$$

where

$$P_U = P_X \Delta_{UY|X}, \quad P_Y = P_X \Delta_{UY|X}, \quad \text{and} \quad Q_{UY} = Q_X \Delta_{UY|X} \ . \tag{3.4}$$

The following lemma asserts the convexity of $\Phi(\cdot)$ and $\Psi(\cdot)$.

LEMMA 5.4. *$\Phi(\Delta_{U|X})$ defined in (3.1) is a convex function in $\Delta_{U|X}$, and $\Psi(\Delta_{UY|X})$ defined in (3.3) is a convex function in $\Delta_{UY|X}$.*

PROOF. For $\alpha \in (0,1)$ and any two conditional distributions $\Delta_{U|X}, \tilde{\Delta}_{U|X}$, let P_U, Q_U be defined as in (3.2), and $\tilde{P}_U, \tilde{Q}_U$ be defined similarly with $\tilde{\Delta}_{U|X}$ replacing $\Delta_{U|X}$. Then

$$\begin{aligned} \Phi(\alpha\Delta_{U|X} + (1-\alpha)\tilde{\Delta}_{U|X}) &= D(\alpha P_U + (1-\alpha)\tilde{P}_U || \alpha Q_U + (1-\alpha)\tilde{Q}_U) \\ &\leq \alpha D(P_U || Q_U) + (1-\alpha) D(\tilde{P}_U || \tilde{Q}_U) \\ &= \alpha\Phi(\Delta_{U|X}) + (1-\alpha)\Phi(\tilde{\Delta}_{U|X}) \ , \end{aligned}$$

where we have made use of the convexity of the divergence. This proves that $\Phi(\Delta_{U|X})$ is a convex function in $\Delta_{U|X}$.

Now let $\alpha \in (0,1)$ and $\Delta_{UY|X}, \tilde{\Delta}_{UY|X}$ be two conditional distributions. We define P_U, P_Y, Q_{UY} as in (3.4), and $\tilde{P}_U, \tilde{P}_Y, \tilde{Q}_{UY}$ similarly with $\tilde{\Delta}_{UY|X}$ replacing $\Delta_{UY|X}$. Then

$$\begin{aligned} \alpha\Psi(\Delta_{UY|X}) + (1-\alpha)\Psi(\tilde{\Delta}_{UY|X}) &= \alpha d(P_U, P_Y || Q_{UY}) + (1-\alpha) d(\tilde{P}_U, \tilde{P}_Y || \tilde{Q}_{UY}) \\ &= \alpha D(P_{UY}^{(1)} || Q_{UY}) + (1-\alpha) D(P_{UY}^{(2)} || \tilde{Q}_{UY}) \end{aligned}$$

for some $P^{(1)}_{UY}$ and $P^{(2)}_{UY}$, where $P^{(1)}_{UY}$ has marginals P_U and P_Y, and $P^{(2)}_{UY}$ has marginals $\tilde{P}_U$ and $\tilde{P}_Y$. By convexity of the divergence functional and the definition of $d(\cdot,\cdot||\cdot)$, we obtain

$$\begin{aligned}
&\alpha\Psi(\Delta_{UY|X}) + (1-\alpha)\Psi(\tilde{\Delta}_{UY|X}) \\
&\qquad \geq\ D(\alpha P^{(1)}_{UY} + (1-\alpha)P^{(2)}_{UY}||\alpha Q_{UY} + (1-\alpha)\tilde{Q}_{UY}) \\
&\qquad \geq\ d(\alpha P_U + (1-\alpha)\tilde{P}_U, \alpha P_Y + (1-\alpha)\tilde{P}_Y||\alpha Q_{UY} + (1-\alpha)\tilde{Q}_{UY}) \\
&\qquad =\ \Psi(\alpha\Delta_{UY|X} + (1-\alpha)\tilde{\Delta}_{UY|X})\ .
\end{aligned}$$

This proves that $\Psi(\Delta_{UY|X})$ is a convex function in $\Delta_{UY|X}$. $\triangle$

In what follows we assume that Π and Λ are partitions of $\mathcal{X}$. We denote by $\Pi \vee \Lambda$ the coarsest common refinement of Π and Λ. We use $P|_\Pi$ to denote the restriction of P_X on Π.

THEOREM 5.4. *Assume all sensors use the same encoder.*

(i) For all $\epsilon \in (0,1)$, $|\mathcal{U}| \leq |\mathcal{X}|$, *if* Π *is a partition of* $\mathcal{X}$, *then*

$$\theta^{(1)}(|\mathcal{U}|,\epsilon) = \max_{\substack{\Pi: \\ P_U = P|_\Pi,\ Q_U = Q|_\Pi}} D(P_U||Q_U)\ .$$

(ii) If $\mathcal{U}$, $\mathcal{Y}$ *are binary sets and* Π, Λ *are partitions of* $\mathcal{X}$, *then for all* $\epsilon \in (0,1)$

$$\theta^{(2)}(\epsilon) = \max_{\substack{\Pi,\Lambda: \\ P_{UY} = P|_{\Pi\vee\Lambda},\ Q_{UY} = Q|_{\Pi\vee\Lambda}}} D(P_{UY}||Q_{UY})\ ,$$

(iii) If, in addition to (ii), $D(P_X||Q_X) < \infty$, *then for all* $\epsilon \in (0,1)$

$$\theta^{(3)}(\epsilon) = \theta^{(4)}(\epsilon) = \max_{\substack{\Pi,\Lambda: \\ P_{UY} = P|_{\Pi\vee\Lambda},\ Q_{UY} = Q|_{\Pi\vee\Lambda}}} d(P_U, P_Y||Q_{UY})\ .$$

PROOF. From Theorem 5.1 and the definition of $\Phi(\cdot)$ we have

$$\theta^{(1)}(|\mathcal{U}|,\epsilon) = \sup_{\Delta_{U|X}} \Phi(\Delta_{U|X}) \ . \tag{3.5}$$

Observe, however, that any distribution $\Delta_{U|X}$ can be written as a convex combination of at most $|\mathcal{U}|^{|\mathcal{X}|}$ extremal distributions $\{\Delta_{U_i|X_i}\}$, which are such that $\Delta_{U_i|X_i}(u|x) = 0$ or 1. Thus if $M = |\mathcal{U}|^{|\mathcal{X}|}$, we can write

$$\Delta_{U|X} = \sum_{i=1}^{M} \alpha_i \Delta_{U_i|X_i} \ ,$$

where $\alpha \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$. Substituting in (3.5) and making use of the convexity of $\Phi(\cdot)$, we obtain

$$\begin{aligned} \theta^{(1)}(|\mathcal{U}|,\epsilon) &= \sup_{\Delta_{U|X}} \Phi(\Delta_{U|X}) = \sup_{\{\alpha_i\}_{i=1}^{M}} \Phi(\sum_{i=1}^{M} \alpha_i \Delta_{U_i|X_i}) \\ &\leq \sup_{\{\alpha_i\}_{i=1}^{M}} \sum_{i=1}^{M} \alpha_i \Phi(\Delta_{U_i|X_i}) \leq \max_{1\leq i\leq M} \Phi(\Delta_{U_i|X_i}) \ . \end{aligned}$$

The reverse inequality is obviously true. This proves the first statment of the theorem. The remaining statments can be proven in a similar way. $\triangle$

5.4. Extensions and concluding remarks.

In the previous section we considered the situation in which all sensors use the same encoder, and showed that no loss of optimality resulted from using deterministic encoders. In this section we consider a more general situation, in which the sensors are allowed to use different encoders. We show that in this case the error exponents of Systems 1 and 2 will still be given by the corresponding expressions in Theorem 5.4, and thus the sensors can use the same deterministic encoder without loss of optimality. We have not yet succeeded in proving an analogous result for Systems 3 and 4. Our main result is thus:

THEOREM 5.5. *If the uniformity constraint on the local encoders are removed, then the error exponents for Systems 1 and 2 are still given by the corresponding expressions in Theorem 5.4.*

PROOF. We give the proof for System 1. Let $\mathcal{U}$ be a set of cardinality not greater than $|\mathcal{X}|$. If $D(P_X||Q_X) = \infty$, then there exists $\bar{x} \in \mathcal{X}$ such that $P_X(\bar{x}) > 0$ and $Q_X(\bar{x}) = 0$. Let U be a message on $\mathcal{U} = \{u_1, \ldots, u_{|\mathcal{U}|}\}$ generated by X via the encoder $\Delta_{U|X}$, where

$$\Delta_{U|X}(u|\bar{x}) = \begin{cases} 1, & \text{if } u = u_1 \ ; \\ 0, & \text{otherwise} \ ; \end{cases}$$

and for any $x \neq \bar{x}$

$$\Delta_{U|X}(u|x) = \begin{cases} 1, & \text{if } u = u_2 \ ; \\ 0, & \text{otherwise} \ . \end{cases}$$

Set the acceptance region $\mathcal{A}_N = T_{U,\eta}$. This yields

$$P_U^N(\mathcal{A}_N) \geq 1 - \epsilon \ , \qquad Q_U^N(\mathcal{A}_N) = 0 \ .$$

Thus $\theta^{(1)}(|\mathcal{U}|, \epsilon) = \infty$. Hence it suffices to prove the theorem under the assumption that $D(P_X||Q_X) < \infty$.

Direct part. Fix any conditional distribution $\Delta_{U|X}$ and let all sensors use the same encoder, i.e., $\Delta_{U_i|X_i} = \Delta_{U|X}$ for all $1 \leq i \leq N$. We obtain the same lower bound as in Theorem 5.1, namely

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) \ \geq \ \sup_{\Delta_{U|X}} D(P_U||Q_U) \ \geq \ \max_{\substack{\Pi: \\ P_U = P|_\Pi, \ Q_U = Q|_\Pi}} D(P_U||Q_U) \ .$$

Converse part. In System 1 assume that each sensor S_i uses an arbitrary encoder $\Delta_{U_i|X_i}$. For all $u_i \in \mathcal{U}$, $i \in \{1, \ldots, N\}$, define

$$P_{U_i} \stackrel{\text{def}}{=} P_X \Delta_{U_i|X_i} \qquad \text{and} \qquad Q_{U_i} \stackrel{\text{def}}{=} Q_X \Delta_{U_i|X_i} \ .$$

It is clear that the U_i's are independent under both hypotheses. Let $\mathcal{A}_N \subset \mathcal{U}^N$ be any acceptance region satisfying the constraint

$$P_{U^N}(\mathcal{A}_N) \geq 1 - \epsilon \, .$$

Define, for $\eta > 0$ arbitrary, the set

$$T_\eta^N \stackrel{\text{def}}{=} \{u^N \in \mathcal{U}^N : \left| \log \frac{P_{U^N}(u^N)}{Q_{U^N}(u^N)} - \sum_{i=1}^N D(P_{U_i} || Q_{U_i}) \right| \leq N\eta \} \, .$$

If $E_P(\cdot)$ and $\mathrm{Var}_P(\cdot)$ respectively denote expectation and variance under P, then

$$\begin{aligned} E_P \log \frac{P_{U^N}(U^N)}{Q_{U^N}(U^N)} &= \sum_{i=1}^N E_P \log \frac{P_{U_i}(U_i)}{Q_{U_i}(U_i)} \\ &= \sum_{i=1}^N D(P_{U_i} || Q_{U_i}) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}_P \log \frac{P_{U^N}(U^N)}{Q_{U^N}(U^N)} &= \sum_{i=1}^N \mathrm{Var}_P \log \frac{P_{U_i}(U_i)}{Q_{U_i}(U_i)} \\ &\leq N \sup_{\Delta_{U|X}} \mathrm{Var}_P \log \frac{P_U(U)}{Q_U(U)} = N\sigma^2 \, , \end{aligned}$$

where $\sigma^2 = \sup_{\Delta_{U|X}} \mathrm{Var}_P \log \frac{P_U(U)}{Q_U(U)}$, and $\sigma^2 < \infty$ by the result of Appendix B. We have from Chebyshev's inequality that

$$\begin{aligned} P_{U^N}(T_\eta^c) &= P_{U^N}\{ \left| \log \frac{P_{U^N}(U^N)}{Q_{U^N}(U^N)} - \sum_{i=1}^N D(P_{U_i} || Q_{U_i}) \right| > N\eta \} \\ &\leq \frac{1}{N^2\eta^2} \mathrm{Var}_P \log \frac{P_{U^N}(U^N)}{Q_{U^N}(U^N)} \leq \frac{\sigma^2}{N\eta^2} \, . \end{aligned}$$

This yields

$$P_{U^N}(\mathcal{A}_N \cap T_\eta) \geq \frac{1-\epsilon}{2}$$

for N sufficiently large. We can estimate the type II error rate as follows:

$$\begin{aligned} Q_{U^N}(\mathcal{A}_N) &\geq Q_{U^N}(\mathcal{A}_N \cap T_\eta) = \sum_{u^N \in \mathcal{A}_N \cap T_\eta^N} Q_{U^N}(u^N) \\ &\geq \sum_{u^N \in \mathcal{A}_N \cap T_\eta^N} P_{U^N}(u^N) \exp\Big[-\sum_{i=1}^{N} D(P_{U_i}||Q_{U_i}) - N\eta\Big] \\ &\geq \frac{1-\epsilon}{2} \exp\Big[-\sum_{i=1}^{N} D(P_{U_i}||Q_{U_i}) - N\eta\Big] \ . \end{aligned}$$

It follows that

$$\theta^{(1)}(|\mathcal{U}|, \epsilon) \leq \frac{1}{N}\sum_{i=1}^{N} D(P_{U_i}||Q_{U_i}) + \eta \leq \sup_{\Delta_{U|X}} D(P_U||Q_U) + \eta \ .$$

Combining the direct and converse parts we obtain the desired conclusion for System 1. For System 2, the proof is similar and is omitted. $\triangle$

The last result of Theorem 5.5 bears some resemblance to that obtained by Tsitsiklis in [23] with the following important differences:

(i) The decision rules in [23] were restricted to be deterministic, whereas we consider the wider class of behavioral rules.

(ii) The optimization in [23] was based on minimizing the overall probability of error, whereas we employ the classical Neyman-Pearson criterion.

(iii) The space of observations in [23] was infinite, whereas here only finite spaces are treated.

We should emphasize that the results of this chapter will still be valid if we assume that the space of observations $\mathcal{X}$ is infinite, but restrict the local encoders to be deterministic and the message alphabets to be finite.

We have been unable to extend Theorem 5.5 so as to include Systems 3 and 4. This is because of the apparent necessity of the positivity assumption on the alternative distribution in Theorem 2.3.

As a final remark, in studying System 2, we assumed that the fusion center transmits a binary feedback message $V \in \mathcal{V}$ ($|\mathcal{V}| = 2$) to all sensors and showed that this system does not offer any improvement over System 1. This is actually true for any finite alphabet $\mathcal{V}$, provided $|\mathcal{V}|$ is fixed in N.

CHAPTER 6

MULTITERMINAL HYPOTHESIS TESTING WITH TIME DEPENDENT OBSERVATIONS

6.1. Introduction

So far the observations have been assumed to be the output of a memoryless multiple source. This assumption is not always justified in practice, since most information sources possess memory. In this chapter we relax this assumption by assuming that the data are available from the simplest example of memory sources, the Markov source.

Once more we are given two hypotheses H_0 and H_1, one of which is true. The system consists of two sensors S_X and S_Y linked to a fusion center (or detector) which decides on the true hypothesis. The sensors S_X and S_Y observe the respective components of the random sequence $\{(X_i, Y_i)\}_{i=1}^n$ ($X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets) and encode their observations using a maximum of $nR_X(n)$ and $nR_Y(n)$ bits, respectively. The detector accepts H_0 if the received sequence lies in an acceptance region $\mathcal{A}_n \in (\mathcal{X} \times \mathcal{Y})^n$. This acceptance region is chosen so as to minimize the type II error rate subject to a fixed upper bound ϵ on the type I error rate.

We assume that the X-data are compressed into a fixed number of codewords and the Y-data are transmited uncompressed, i.e., $nR_X(n) = \text{constant}$ and $R_Y(n) = \infty$. An asymptotically optimal acceptance region for the corresponding memoryless system has been determined in Chapter 2, and the error

exponent was shown to be equal to

$$\min_{\substack{\tilde{P}_{XY}: \\ \tilde{P}_X = P_X,\ \tilde{P}_Y = P_Y}} D(\tilde{P}_{XY} || Q_{XY}) \ ,$$

where P_{XY} (resp. Q_{XY}) denotes the distribution of (X_i, Y_i) under H_0 (resp. H_1).

If, under both hypotheses, the random process $\{(X_i, Y_i)\}_{i=-\infty}^{\infty}$ forms a 1st order Markov chain whose transition matrix is irreducible, and there is no compression on either X or Y-data (i.e., $R_X(n) = R_Y(n) = \infty$), then the problem is considerably simplified because the detector knows the observed sequence $\{(X_i, Y_i)\}_{i=1}^{n}$ precisely. In this case the optimal acceptance region can be found in [35,36] and the resulting minimum type II error rate is given by

$$D(W||V|\pi_W) \ ,$$

where $D(\cdot||\cdot|\cdot)$ denotes informational divergence rate, W (resp. V) the transition matrix of the Markov process $\{(X_i, Y_i)\}$ under H_0 (resp. H_1), and π_W the stationary distribution of the Markov process under the null hypothesis.

6.2. Problem statement and preliminaries

(a) General notation. We assume that we have a double-sided process $\{(X_i, Y_i),\ i \in \mathbf{Z}\}$, and X_i, Y_i take values in the finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. We also let $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{Y}$ and $Z_i \stackrel{\text{def}}{=} (X_i, Y_i)$. The sensors S_X and S_Y observe the finite random sequences $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, respectively.

If $i \leq j$, then $Z_i^j = (XY)_i^j$ will denote the random sequence $(Z_i, \ldots, Z_j) = ((X_i, Y_i), \ldots, (X_j, Y_j))$ on $\mathcal{Z}^{j-i+1}$, and $z_i^j = (xy)_i^j$ will denote the corresponding nonrandom element $(z_i, \ldots, z_j) = ((x_i, y_i), \ldots, (x_j, y_j))$. Similar notation will also be adopted using X_i^j, x_i^j, Y_i^j, and y_i^j.

We will use the symbols W_k and V_k to denote transition matrices for (finite order) Markov processes. The subscript k in W_k and V_k indicates that the order of the process equals $k-1$. 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 following a given sequence $z_1^{k-1} = (z_1, \ldots, z_{k-1})$.

Given a stochastic matrix W_k (resp. V_k) on $\mathcal{Z}^k$ which induces a unique stationary distribution π_W (resp. π_V) on $\mathcal{Z}^{k-1}$, we define a stationary $(k-1)$th order Markov measure $P = \pi_W \diamond W_k$ (resp. $Q = \pi_V \diamond V_k$) on the Borel field of $\mathcal{Z}^{\mathbf{Z}}$ by means of the following finite-dimensional distributions:

$$P(z_{n_1}^{n_2}) = \pi_W(z_{n_1}^{n_1+k-2}) \prod_{i=n_1}^{n_2-k+1} W_k(z_{i+k-1}|z_i^{i+k-2}) ,$$

$$Q(z_{n_1}^{n_2}) = \pi_V(z_{n_1}^{n_1+k-2}) \prod_{i=n_1}^{n_2-k+1} V_k(z_{i+k-1}|z_i^{i+k-2}) ,$$

for all $n_1 \leq n_2$ and $z_{n_1}^{n_2} \in \mathcal{Z}^{n_2-n_1+1}$. The entropy rate of the corresponding $(k-1)$th order Markov process (π_W, W_k) is defined by

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

where $P = \pi_W \diamond W_k$. If V_l is a stochastic matrix on $\mathcal{Z}^l$ with $l \leq k$, we define the divergence rate $D(W_k||V_l|\pi_W)$ by

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

The compression of X_1^n is effected by the encoder f_n, where

$$f_n : \mathcal{X}^n \mapsto \{1, \ldots, M\} ,$$

and the codebook size M is constrained by $M \geq 2$.

The corresponding detector is represented by the function

$$\phi_n : \{1, \ldots, M\} \times \mathcal{Y}^n \mapsto \{0, 1\} \ ,$$

where the output 0 signifies the acceptance of the null hypothesis H_0, and 1 its rejection. This induces a partition of the original (i.e., non-compressed) sample space $(\mathcal{X} \times \mathcal{Y})^n$ into an *acceptance* region

$$\mathcal{A}_n \stackrel{\text{def}}{=} \{(xy)_1^n \in (\mathcal{X} \times \mathcal{Y})^n : \phi_n(f_n(x_1^n), y_1^n) = 0\}$$

and a rejection region $\mathcal{A}_n^c$. By nature of the encoding process, the acceptance region can be decomposed into M rectangles $C_i \times F_i$ in $(\mathcal{X} \times \mathcal{Y})^n$ that possess disjoint projections C_i on $\mathcal{X}^n$, where

$$C_i \times F_i \stackrel{\text{def}}{=} \{(xy)_1^n \in (\mathcal{X} \times \mathcal{Y})^n : x_1^n \in C_i, \ y_1^n \in F_i\} \ .$$

More precisely, if for every $1 \leq i \leq M$ we define

$$C_i \ = \ \{x_1^n \in \mathcal{X}^n : f_n(x_1^n) = i\} \qquad \text{and} \qquad F_i \ = \ \{y_1^n \in \mathcal{Y}^n : \phi_n(i, y_1^n) = 0\} \ ,$$

then we can write

$$\mathcal{A}_n \ = \ \bigcup_{i=1}^{M} C_i \times F_i \ , \qquad \text{where} \qquad (\forall i \neq j) \quad C_i \cap C_j = \emptyset \ , \tag{2.1}$$

(b) The hypothesis testing problem. In the above framework, the problem of testing $H_0 : W = W_2$ versus $H_1 : V = V_2$ can be formulated as follows: for a given level $\epsilon \in (0, 1)$, minimize $Q(\mathcal{A}_n)$ (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 (2.1).

The resulting *minimum* probability of type II error is denoted by $\beta_n(M, \epsilon)$, and the associated error exponent is given by

$$\theta(M, \epsilon) \stackrel{\text{def}}{=} -\lim_n \frac{1}{n} \beta_n(M, \epsilon) \ ,$$

provided the limit on the right-hand side exists.

We restrict ourselves to the study of the asymptotic behavior of $\beta_n(M, \epsilon)$ as n approaches infinity.

Under the conditions:

(C1) the process $\{(X_i, Y_i),\ i \in \mathbf{Z}\}$ is stationary ergodic (1st order) Markov under both hypotheses;

(C2) W is irreducible and aperiodic; and

(C3) $V > 0$,

we show that for every $M \geq 2$, $\epsilon \in (0, 1)$, the error exponent $\theta(M, \epsilon)$ is given by the infimum of the quantity

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

over all stationary distributions $\tilde{P}$ on the Borel field of $(\mathcal{X} \times \mathcal{Y})^{\mathbf{Z}}$ that satisfy for all $n \geq 0$ the conditions $\tilde{P}(x_{-n}^0) = P(x_{-n}^0)$ and $\tilde{P}(y_{-n}^0) = P(y_{-n}^0)$.

(c) Typical sequences. Our proofs in this chapter rely on a broader concept of a typicality than the one employed in the previous chapters.

The *kth order type* of a sequence $z_1^n \in \mathcal{Z}^n$ is the distribution $\lambda_{z_1^n}^{(k)}$, or simply $\lambda_z^{(k)}$, on $\mathcal{Z}^k$ defined by the relationship

$$(\forall a_1^k \in \mathcal{Z}^k) \qquad \lambda_z^{(k)}(a_1^k) \stackrel{\text{def}}{=} \frac{1}{n} \left|\{i \in \{1,\ldots,n\}\colon z_i^{k+i-1} = a_1^k\}\right| ,$$

where the circular convention $z_{n+j} = z_j,\ j \in \{1,\ldots,k-1\}$, is adopted. Thus the 1st order type is the ordinary type encountered earlier, while the 2nd order type is the "Markov type" introduced by Davisson *et al.* [37]. The circular convention used in the above definition ensures that

$$\sum_{b\in\mathcal{Z}} \lambda_z^{(k)}(b, a_1, \ldots, a_{k-1}) \;=\; \sum_{b\in\mathcal{Z}} \lambda_z^{(k)}(a_1, \ldots, a_{k-1}, b) \;=\; \lambda_z^{(k-1)}(a_1^{k-1}) \qquad (2.2)$$

for all $a_1^{k-1} \in \mathcal{Z}^{k-1}$. As a consequence of the above identity, the marginals of $\lambda_z^{(k)}$ are "stationary," i.e., translation invariant within the index set $\{1,\ldots,k\}$.

The set of all kth order types is denoted by $\mathcal{P}_n^{(k)}(\mathcal{Z})$. It is easy to see that the cardinality of this set is at most $(n+1)^{|\mathcal{Z}|^k}$.

If $\hat{P}_k \in \mathcal{P}_n^{(k)}(\mathcal{Z})$, we define $\hat{T}_Z^{(k)} \subset \mathcal{Z}^n$ to be the set of sequences $z_1^n \in \mathcal{Z}^n$ that have kth order type $\lambda_z^{(k)} = \hat{P}_k$. Let the distribution $\hat{\pi}_W$ and the stochastic matrix $\hat{W}_k$ be defined by

$$(\forall z_1^{k-1} \in \mathcal{Z}^{k-1}) \qquad \hat{\pi}_W(z_1^{k-1}) \stackrel{\text{def}}{=} \sum_{z_k} \hat{P}_k(z_1^k) \qquad (2.3)$$

and

$$(\forall z_1^k \in \mathcal{Z}^k) \qquad \hat{W}_k(z_k|z_1^{k-1}) \stackrel{\text{def}}{=} \begin{cases} \hat{P}_k(z_1^k)/\hat{\pi}_W(z_1^{k-1}), & \text{if } \hat{\pi}_W(z_1^{k-1}) > 0\ ; \\ 1/|\mathcal{Z}|, & \text{otherwise}\ . \end{cases} \qquad (2.4)$$

It is easy to see that $\hat{\pi}_W$ defined previously is a stationary distribution for $\hat{W}_k$. The circular convention of kth order type guarantees that $\hat{W}_k$ defined in (2.4) is

irreducible, and hence $\hat{\pi}_W$ is the unique stationary distribution. We also have $\hat{P}_k = \hat{\pi}_W \diamond \hat{W}_k$.

LEMMA 6.1. *Let $\hat{P}_k \in \mathcal{P}_n^{(k)}(\mathcal{Z})$ and $\hat{\pi}_W$, $\hat{W}_k$ be as above. Then*

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

PROOF. Davisson *et al.* [37] have proved the above lemma for $k = 2$. Following the same procedure we can show that the lemma is also true for $k > 2$.

$\triangle$

LEMMA 6.2. *Let $\hat{P}_k \in \mathcal{P}_n^{(k)}(\mathcal{Z})$ and $\hat{\pi}_W$, $\hat{W}_k$ be as in (2.3), (2.4). For any $l \leq k$, if $V_l > 0$, then*

$$\alpha_l n^{-|\mathcal{Z}|^{k-1}}(n+1)^{-|\mathcal{Z}|^k} \exp[-nD(\hat{W}_k||V_l|\hat{\pi}_W)] \ \leq \ Q(\hat{T}_Z^{(k)}) \ \leq$$
$$\leq \frac{|\mathcal{Z}|^{k-1}}{\rho_l^{l-1}} \exp[-nD(\hat{W}_k||V_l|\hat{\pi}_W)] \ ,$$

where $\rho_l = \min_{z_1^l \in \mathcal{Z}^l} V_l(z_l|z_1^{l-1})$ and $\alpha_l = \min_{z_1^{l-1} \in \mathcal{Z}^{l-1}} \pi_V(z_1^{l-1})$.

PROOF. Let $z_1^n \in \hat{T}_Z^{(k)}$. From the convention $z_{n+j} = z_j$, $j \in \{1, \ldots, k-1\}$, and the fact that a $(l-1)$th order Markov chain is also kth order Markov for $l \leq k$, we obtain

$$\begin{aligned} Q(z_1^n) &= Q(z_1^{k-1}) \prod_{i=1}^{n-k+1} V_l(z_{i+k-1}|z_i^{i+k-2}) \\ &= c(z_1^n, l) \prod_{i=1}^{n} V_l(z_{i+k-1}|z_i^{i+k-2}) \\ &= c(z_1^n, l) \prod_{z_1^k \in \mathcal{Z}^k} V_l(z_k|z_1^{k-1})^{n\hat{P}_k(z_1^k)} \\ &= c(z_1^n, l) \prod_{z_1^k \in \mathcal{Z}^k} V_l(z_k|z_{k-l+1}^{k-1})^{n\hat{P}_k(z_1^k)} \ , \end{aligned}$$

where

$$c(z_1^n, l) = Q(z_1^{k-1}) / \prod_{i=n-k+2}^{n} V_l(z_{i+k-1}|z_i^{i+k-2})$$
$$= \pi_V(z_1^{l-1}) / \prod_{i=n-l+2}^{n} V_l(z_{i+l-1}|z_i^{i+l-2}) .$$

Using the fact that $c(z_1^n, l)$ is neither less than α_l nor greater than $\rho_l^{-(l-1)}$ together with Lemma 6.1 completes the proof. $\triangle$

Given a process $\{Z_i,\ i \in \mathbf{Z}\}$ with probability measure P, we define, for any $\eta > 0$, the set of kth order (P, η)-typical sequences $T_{Z,\eta}^{(k)} \subset \mathcal{Z}^n$ as

$$T_{Z,\eta}^{(k)} \stackrel{\text{def}}{=} \{z_1^n \in \mathcal{Z}^n\colon \sup_{a_1^k} |\lambda_z^{(k)}(a_1^k) - P(a_1^k)| \le \eta\} .$$

LEMMA 6.3. *If $\{Z_i,\ i \in \mathbf{Z}\}$ is a stationary ergodic process with distribution P, then for any $k \ge 1$ and $\eta > 0$, there exists a sequence $\{\xi_n\}_{n=1}^{\infty}$ with $\xi_n \to 0$ such that*

$$P(T_{Z,\eta}^{(k)}) \ge 1 - \xi_n .$$

PROOF. The pointwise ergodic theorem [38] implies that for any $k \ge 1$,

$$P\{z_{-\infty}^{\infty} \in \mathcal{Z}^{\mathbf{Z}}\colon (\forall a_1^k \in \mathcal{Z}^k) \quad \lim_n \lambda_{z_1^n}^{(k)}(a_1^k) = P(a_1^k)\} = 1 .$$

Since convergence almost everywhere implies convergence in probability, we obtain for arbitrary $\eta > 0$,

$$\lim_n P\{z_1^n \in \mathcal{Z}^n\colon (\forall a_1^k \in \mathcal{Z}^k) \quad |\lambda_{z_1^n}^{(k)}(a_1^k) - P(a_1^k)| \le \eta\} = 1 . \qquad \triangle$$

6.3. The main results

Let $W = W_2$ be irreducible and aperiodic (so that $\pi_W > 0$), and $P = \pi_W \diamond W$. For all $k \geq 2$ we define the following linear subspaces of distributions on $\mathcal{Z}^{\mathbf{Z}}$.

$$\mathcal{L}^k \stackrel{\text{def}}{=} \{\tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbf{Z}}:$$
$$\tilde{P}(x_{-k+1}^0) = P(x_{-k+1}^0),\ \tilde{P}(y_{-k+1}^0) = P(y_{-k+1}^0)\} \ . \qquad (3.1)$$

It is obvious that $\mathcal{L}^{k+1} \subset \mathcal{L}^k$ and $\mathcal{L}^k \searrow \mathcal{L}$, where

$$\mathcal{L} \stackrel{\text{def}}{=} \{\tilde{P} \text{ stationary on } \mathcal{Z}^{\mathbf{Z}}:$$
$$(\forall n \geq 0) \quad \tilde{P}(x_{-n}^0) = P(x_{-n}^0),\ \tilde{P}(y_{-n}^0) = P(y_{-n}^0)\} \ . \qquad (3.2)$$

Our aim in this section is to show that the error exponent $\theta(M, \epsilon)$ for the hypothesis testing problem formulated in Section 6.2.(b) is given by

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

The positive result is as follows.

THEOREM 6.1. *If we define for all $k \geq 2$*

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

then for all $\epsilon \in (0,1)$ and $M \geq 2$, the error exponent for the hypothesis testing problem formulated in Section 6.2.(b) satisfies

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

PROOF. Fix $k \geq 2$ and $\eta > 0$. Since W is irreducible and aperiodic, the distribution P on $\mathcal{Z}^{\mathbf{Z}}$ is strongly mixing. This implies that the marginals of P on $\mathcal{X}^{\mathbf{Z}}$ and $\mathcal{Y}^{\mathbf{Z}}$ are strongly mixing as well, and therefore ergodic. Let $T_{X,\eta}^{(k)} \subset \mathcal{X}^n$, $T_{Y,\eta}^{(k)} \subset \mathcal{Y}^n$ be the corresponding sets of kth order typical sequences. Consider the following acceptance region:

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

By Lemma 6.3, there exists a sequence $\{\xi_n\}$ with $\xi_n \to 0$ such that

$$P(\mathcal{A}_n) \;\geq\; P(T_{X,\eta}^{(k)}) + P(T_{Y,\eta}^{(k)}) - 1 \;\geq\; 1 - 2\xi_n .$$

Hence, for n large enough,

$$P(\mathcal{A}_n) \;\geq\; 1 - \epsilon .$$

We can write

$$Q(\mathcal{A}_n) \;=\; Q\Big(\bigcup_{\hat{P}_k \in \Phi^k} \hat{T}_{(XY)}^{(k)}\Big) ,$$

where

$$\Phi^k = \{\hat{P} \in \mathcal{P}_n^{(k)}(\mathcal{X} \times \mathcal{Y})\colon \sup_{x_1^k} |\hat{P}(x_1^k) - P(x_1^k)| \leq \eta,\ \sup_{y_1^k} |\hat{P}(y_1^k) - P(y_1^k)| \leq \eta\} .$$

Using Lemma 6.2 we obtain

$$\begin{aligned} \beta_n(M, \epsilon) \;&\leq\; Q(\mathcal{A}_n) \\ &\leq\; \exp[-n(\min_{\hat{P}_k \in \Phi^k} D(\hat{W}_k || V | \hat{\pi}_W) - \delta_n)] . \end{aligned}$$

Here $\hat{P}_k = \hat{\pi}_W \diamond \hat{W}_k$, where $\hat{\pi}_W$ and $\hat{W}_k$ are given by (2.3) and (2.4), respectively, and

$$\delta_n = |\mathcal{Z}|^k \frac{\log(n+1)}{n} + \frac{1}{n} \log \frac{|\mathcal{Z}|^{k-1}}{\rho} \to 0 ,$$

where $\rho = \min_{z_{-1}^0 \in \mathcal{Z}^2} V\{z_0|z_{-1}\}$. Continuity of the divergence functional enables us to approximate the above minimum over Φ^k by one over $\mathcal{S}^k$, where $\mathcal{S}^k$ is the class of all stationary distributions on $(\mathcal{X} \times \mathcal{Y})^k$ that agree with P on both $\mathcal{X}^k$ and $\mathcal{Y}^k$. Indeed, $\Phi^k \subset \mathcal{S}^k$, and any distribution $\breve{P} \in \mathcal{S}^k$ can be approximated by a positive type $\hat{P}_k \in \Phi^k$ with irreducible transition matrix $\hat{W}_k$. Thus for sufficiently small $\eta > 0$, we have

$$\beta_n(M,\epsilon) \;\leq\; \exp[-n(\min_{\breve{P}\in\mathcal{S}^k} \sum_{(xy)_1^k} \breve{P}\{(xy)_1^k\} \log \frac{\breve{P}\{(xy)_k|(xy)_1^{k-1}\}}{V\{(xy)_k|(xy)_{k-1}\}} - \delta_n - \nu(\eta))] \;,$$

where $\nu(\eta) \to 0$ as $\eta \to 0$. Thus,

$$\begin{aligned}
\theta(M,\epsilon) \;&\geq\; \liminf_n(-\frac{1}{n}\log\beta_n(M,\epsilon)) \\
&\geq\; \min_{\breve{P}\in\mathcal{S}^k} \sum_{(xy)_1^k} \breve{P}\{(xy)_1^k\} \log \frac{\breve{P}\{(xy)_k|(xy)_1^{k-1}\}}{V\{(xy)_k|(xy)_{k-1}\}} \\
&=\; \min_{\breve{P}\in\mathcal{L}^k} E_{\breve{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-k+1}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} \;,
\end{aligned}$$

where $\mathcal{L}^k$ is the infinite-dimensional counterpart of $\mathcal{S}^k$ defined in (3.1). Thus we have shown that $\theta(M,\epsilon) \geq \sup_k D^k$.

Next we show that D^k is monotone increasing in k. Using the log-sum inequality, we can write

$$\begin{aligned}
D^{k+1} \;&=\; \sum_{(xy)_{-k}^0} \mu_{k+1}\{(xy)_{-k}^0\} \log \frac{\mu_{k+1}\{(xy)_{-k}^0\}}{\mu_{k+1}\{(xy)_{-k}^{-1}\}V\{(xy)_0|(xy)_{-1}\}} \\
&\geq\; \sum_{(xy)_{-k+1}^0} \mu_{k+1}\{(xy)_{-k+1}^0\} \log \frac{\mu_{k+1}\{(xy)_{-k+1}^0\}}{\mu_{k+1}\{(xy)_{-k+1}^{-1}\}V\{(xy)_0|(xy)_{-1}\}} \\
&=\; E_{\mu_{k+1}} \log \frac{\mu_{k+1}\{(XY)_0|(XY)_{-k+1}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} \;.
\end{aligned}$$

With the aid of the above inequality, the fact that $\mu_{k+1} \in \mathcal{L}^{k+1} \subset \mathcal{L}^k$, and the definition of D^k, we obtain

$$D^{k+1} \geq D^k \;.$$

Thus D^k is monotone increasing in k and we conclude that $\sup_k D^k = \lim_k D^k$.

In order to complete the proof we need to show that

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

We begin by noting that we have an infinite sequence $\{\mu_k,\ k \in \mathbf{N}\}$ of measures on $(\mathcal{X} \times \mathcal{Y})^{\mathbf{Z}}$ with the constraints $\mu_k(x_{-k+1}^0) = P(x_{-k+1}^0)$ and $\mu_k(y_{-k+1}^0) = P(y_{-k+1}^0)$ for every $x_{-k+1}^0 \in \mathcal{X}^k$ and $y_{-k+1}^0 \in \mathcal{Y}^k$. Since $\mathcal{Z}$ is finite, the infinite product $\mathcal{Z}^{\mathbf{Z}}$ is compact under the product topology induced by the discrete topology on $\mathcal{Z}$. The class of cylinder sets constitutes a countable base for the product topology, and it is easily verified that $\mathcal{Z}^{\mathbf{Z}}$ is Hausdorff. By the Urysohn metrization theorem, $\mathcal{Z}^{\mathbf{Z}}$ is metrizable. Thus $\{\mu_k\}$ is a sequence of measures on the Borel σ-field of a compact metric space. Invoking Prohorov's theorem [45, p. 315], we conclude that $\{\mu_k\}$ contains a subsequence $\{\mu_{k_i},\ i \in \mathbf{N}\}$ which converges weakly to a measure $\tilde{\mu}$. In particular, since every cylinder $G \subset \mathcal{Z}^{\mathbf{Z}}$ is both open and closed, we have

$$\lim_i \mu_{k_i}(G) \ = \ \tilde{\mu}(G) \ .$$

This implies that $\tilde{\mu}$ lies in the class $\mathcal{L}$ defined in (3.2). Thus to conclude the proof of the theorem 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

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

This can be written as

$$\tilde{D}_n = \sum_{z_{-n}^0} \tilde{\mu}(z_{-n}^0) \log \frac{\tilde{\mu}(z_0|z_{-n}^{-1})}{V(z_0|z_{-1})} = E_{\tilde{\mu}} e_n(Z_{-\infty}^{-1}) ,$$

where

$$e_n(Z_{-\infty}^{-1}) \stackrel{\text{def}}{=} \sum_{z_0} \tilde{\mu}(z_0|Z_{-n}^{-1}) \log \frac{\tilde{\mu}(z_0|Z_{-n}^{-1})}{V(z_0|Z_{-1})} .$$

Lévy's martingale convergence theorem for conditional probabilities [45, p. 478] ensures that for all $z_0 \in \mathcal{Z}$,

$$\tilde{\mu}(z_0|Z_{-n}^{-1}) \to \tilde{\mu}(z_0|Z_{-\infty}^{-1}) \qquad \text{a.e. } (\tilde{\mu}) .$$

Hence $e_n \to e_\infty$ almost everywhere, where

$$e_\infty(Z_{-\infty}^{-1}) \stackrel{\text{def}}{=} \sum_{z_0} \tilde{\mu}(z_0|Z_{-\infty}^{-1}) \log \frac{\tilde{\mu}(z_0|Z_{-\infty}^{-1})}{V(z_0|Z_{-1})} .$$

Since $V > 0$, we have $|e_n| \le \log|\mathcal{Z}| + \log\frac{1}{\rho}$ for all n, and the bounded convergence theorem implies that

$$\tilde{D}_n = \int e_n d\tilde{\mu} \to \int e_\infty d\tilde{\mu} = \tilde{D} .$$

Actually, $\tilde{D}_n \nearrow \tilde{D}$ because the difference $\tilde{D}_{n+1} - \tilde{D}_n$ equals the conditional divergence $E_{\tilde{\mu}} \log \frac{\tilde{\mu}(Z_0|Z_{-n-1}^{-1})}{\tilde{\mu}(Z_0|Z_{-n}^{-1})}$, which is nonnegative.

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

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

Indeed, since $\mu_{k_i}(z_{-n}^0) \xrightarrow{i} \tilde{\mu}(z_{-n}^0)$ for all $z_{-n}^0 \in \mathcal{Z}^{n+1}$, we have for all $n > 0$

$$D_n^{k_i} \xrightarrow{i} \tilde{D}_n \ .$$

Step 3. We show that $D^k \geq D_n^k$ for all $k \geq n+1$. This is immediate from

$$D^k - D_n^k \ = \ E_{\mu_k} \log \frac{\mu_k(Z_0|Z_{-k+1}^{-1})}{\mu_k(Z_0|Z_{-n}^{-1})} \ \geq 0 \ .$$

Step 4. Combining the results of steps 2 and 3 yields

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

Taking the limit as $n \to \infty$ and making use of the result of Step 1, we obtain

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

REMARK 1. The stationary measure $\mu_k \in \mathcal{L}^k$ that achieves D^k can be taken to be $(k-1)$th order Markov. Its ergodicity follows from the fact that it is the Markov I-projection of V on $\mathcal{L}^k$ (cf. Csiszár *et al.* [39]). By virtue of this property, for any $n \geq k-1$, for all nth Markov measures $\tilde{P} \in \mathcal{L}^k$,

$$E_{\tilde{P}} \log \frac{\tilde{P}(Z_0|Z_{-n}^{-1})}{V(Z_0|Z_{-1})} = D^k + E_{\tilde{P}} \log \frac{\tilde{P}(Z_0|Z_{-n}^{-1})}{\mu_k(Z_0|Z_{-k+1}^{-1})}$$

or

$$\sum_{z_{-k+1}^0} (\tilde{P}(z_{-k+1}^0) - \mu_k(z_{-k+1}^0)) \log \frac{\mu_k(z_0|z_{-k+1}^{-1})}{V(z_0|z_{-1})} \ = \ 0 \ .$$

In particular if $\tilde{P} = P$, then $\mu_k(z_0|z_{-k+1}^{-1}) > 0$ whenever $P(z_0|z_{-k+1}^{-1}) > 0$, and μ_k is irreducible and aperiodic.

REMARK 2. We have shown in the proof of the theorem that $D^{k+1} \geq D^k$. We will demonstrate by an example that D^{k+1} may be strictly greater than D^k for all k. Indeed, since $\mu_{k+1} \in \mathcal{L}^k$,

$$D^{k+1} = D^k + E_{\mu_{k+1}} \log \frac{\mu_{k+1}\{(XY)_0|(XY)_{-k}^{-1}\}}{\mu_k\{(XY)_0|(XY)_{-k+1}^{-1}\}} \ . \tag{3.3}$$

In the example, take the alternative stochastic matrix $V\{(xy)_0|(xy)_{-1}\}$ to be equal to $V(x_0|x_{-1}) \times V(y_0|y_{-1})$. We show that in this case the sequence $\{\mu_k,\ k \in \mathbf{N}\}$ is given by

$$\mu_k\{(xy)_{-k+1}^0\} = P(x_{-k+1}^0) \times P(y_{-k+1}^0) \ ,$$

and thus

$$D^k = E_P \log \frac{P(X_0|X_{-k+1}^{-1})}{V(X_0|X_{-1})} + E_P \log \frac{P(Y_0|Y_{-k+1}^{-1})}{V(Y_0|Y_{-1})} \ .$$

Indeed, we can write

$$\begin{aligned}
&D^k - E_P \log \frac{P(X_0|X_{-k+1}^{-1})}{V(X_0|X_{-1})} + E_P \log \frac{P(Y_0|Y_{-k+1}^{-1})}{V(Y_0|Y_{-1})} \\
&\quad = \min_{\tilde{P} \in \mathcal{L}^k} \Big[E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-k+1}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} - E_{\tilde{P}} \log \frac{P(X_0|X_{-k+1}^{-1})}{V(X_0|X_{-1})} \\
&\qquad\qquad - E_{\tilde{P}} \log \frac{P(Y_0|Y_{-k+1}^{-1})}{V(Y_0|Y_{-1})} \Big] \\
&\quad = \min_{\tilde{P} \in \mathcal{L}^k} E_{\tilde{P}} \Big[\log \frac{\tilde{P}\{(XY)_0|(XY)_{-k+1}^{-1}\}}{V\{(XY)_0|(XY)_{-1}\}} - \log \frac{P(X_0|X_{-k+1}^{-1}) \times P(Y_0|Y_{-k+1}^{-1})}{V(X_0|X_{-1}) \times V(Y_0|Y_{-1})} \Big] \\
&\quad = \min_{\tilde{P} \in \mathcal{L}^k} E_{\tilde{P}} \log \frac{\tilde{P}\{(XY)_0|(XY)_{-k+1}^{-1}\}}{P(X_0|X_{-k+1}^{-1}) \times P(Y_0|Y_{-k+1}^{-1})} \ .
\end{aligned}$$

The value of the last minimum is zero, and is achieved when $\tilde{P}\{(xy)_{-k+1}^0\} = P(x_{-k+1}^0) \times P(y_{-k+1}^0)$. Substituting in (3.3) yields

$$D^{k+1} = D^k + I(X_0 \wedge X_{-k}|X_{-k+1}^{-1}) + I(Y_0 \wedge Y_{-k}|Y_{-k+1}^{-1}),$$

where $I(\cdot \wedge \cdot|\cdot)$ is the conditional mutual information under P. This implies that D^{k+1} is strictly greater than D^k if either the process $\{X_i\}$ or $\{Y_i\}$ is not $(k-1)$th order Markov under P.

The following theorem asserts the tightness of the error exponent given in the above theorem.

THEOREM 6.2. *For all $\epsilon \in (0,1)$ and $M \geq 2$, the error exponent for the hypothesis testing problem formulated in Section 6.2.(b) satisfies*

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

PROOF. Let $\tilde{\mu} \in \mathcal{L}$ yield a divergence $\tilde{D}$ which is close in value to the above infimum. Let

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

be the optimal acceptance region, 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 a $j \in \{1,\ldots,M\}$ such that

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

Let $C_n \stackrel{\text{def}}{=} C_n^{(j)}$, $F_n \stackrel{\text{def}}{=} F_n^{(j)}$, and $G_n \stackrel{\text{def}}{=} C_n \times F_n$. Thus, we have for $\lambda \in (0, \frac{1-\epsilon}{M})$

$$P(C_n \times F_n) \geq \lambda .$$

Since $\{Z_i,\ i \in \mathbf{Z}\}$ is a strongly mixing Markov process, it is also weakly Bernoulli and thus finitely determined (cf. Ornstein [40]). For such a process, Lemma 1 in [41] implies the following: if the probability that Z_1^n lies in $G_n \subset \mathcal{Z}^n$ is bounded away from zero for all values of n, i.e.,

$$\Pr(Z_1^n \in G_n) \ \geq \ \lambda ,$$

and if $\hat{Z}_1^n$ is a random vector distributed on G_n according to the conditional distribution induced by the distribution of Z_1^n, then a joint distribution $\mathrm{dist}(Z_1^n, \hat{Z}_1^n)$ exists such that

$$\frac{1}{n} E d_H(Z_1^n, \hat{Z}_1^n) \leq \lambda \ .$$

Here $d_H(\cdot,\cdot)$ denotes Hamming distance (as before), and the expectation is with respect to $\mathrm{dist}(Z_1^n, \hat{Z}_1^n)$. Following the same procedure as in Marton [16], we obtain

$$P(\Gamma^k G_n) \ \geq \ 1 - \frac{n\lambda}{k_n} \ ,$$

where

$$\Gamma^k G_n \stackrel{\mathrm{def}}{=} \{\bar{z}_1^n \in \mathcal{Z}^n : (\exists z_1^n \in G_n) \quad d_H(z_1^n, \bar{z}_1^n) \leq k_n\} \ .$$

If $k_n = \lceil n\sqrt{\lambda} \rceil$, the last inequality becomes

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

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

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

which in turn yields

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

By virtue of the marginal constraints on $\tilde{\mu}$, we have

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

For the sake of simplicity we write

$$z = z_{-\infty}^{\infty} \qquad \text{and} \qquad Z = Z_{-\infty}^{\infty} \ .$$

Let

$$E_n \stackrel{\text{def}}{=} (\Gamma^k C_n \times \Gamma^k F_n) \cap \{z_1^n : \tilde{\mu}(z_1^n) > 0\} \ .$$

$Q(E_n)$ can be estimated as follows:

$$\begin{aligned} Q(E_n) &= \sum_{z_1^n \in E_n} Q(z_1^n) \\ &= \sum_{z_1^n \in E_n} \exp[-n i_n(z)] \tilde{\mu}(z_1^n) \ , \end{aligned}$$

where

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

Making use of Jensen's inequality, we obtain

$$\begin{aligned} Q(E_n) &= \tilde{\mu}(E_n) \sum_{z_1^n \in E_n} \frac{\tilde{\mu}(z_1^n)}{\tilde{\mu}(E_n)} \exp[-n i_n(z)] \\ &\geq \tilde{\mu}(E_n) \exp\Big[-n \sum_{z_1^n \in E_n} \frac{\tilde{\mu}(z_1^n)}{\tilde{\mu}(E_n)} i_n(z) \Big] \\ &\geq (1 - 2\sqrt{\lambda}) \exp\Big[-n \frac{\alpha_n}{\tilde{\mu}(E_n)} \Big] \ , \end{aligned} \tag{3.4}$$

where

$$\alpha_n \stackrel{\text{def}}{=} \sum_{z_1^n \in E_n} \tilde{\mu}(z_1^n) i_n(z) = \int_{E_n} i_n(z) d\tilde{\mu}(z) \ .$$

Since $V > 0$, it follows from a version of the Shannon-McMillam-Breiman theorem [42, 43] that $i_n(Z) \to \tilde{\imath}(Z)$ in the $L^1(\tilde{\mu})$ norm, and $\int \tilde{\imath} d\tilde{\mu} = \tilde{D}$. More precisely,

$$\int |i_n(z) - \tilde{\imath}(z)| d\tilde{\mu}(z) \to 0 \ .$$

Next, we will show that α_n can be approximated by $\tilde{D}$. Indeed, we can write

$$|\alpha_n - \tilde{D}| \ = \ | \int_{E_n} i_n d\tilde{\mu} - \int \tilde{\imath} d\tilde{\mu}|$$

$$
\begin{aligned}
&= \Big| \int_{E_n} i_n d\tilde{\mu} - \int_{E_n} \tilde{i} d\tilde{\mu} - \int_{E_n^c} \tilde{i} d\tilde{\mu} \Big| \\
&\leq \int_{E_n} |i_n - \tilde{i}| d\tilde{\mu} + \int_{E_n^c} |\tilde{i}| d\tilde{\mu} \\
&\leq \int |i_n - \tilde{i}| d\tilde{\mu} + \int_{E_n^c} |\tilde{i}| d\tilde{\mu} \ .
\end{aligned}
$$

Given $\xi > 0$, the first integral will be less than $\xi/2$ for n sufficiently large by virtue of $L^1(\tilde{\mu})$ convergence. To upper bound the second integral, we use Prop. 13 in [44] which states that given an integrable nonnegative function $f(z)$ and $\xi > 0$ there exists a $\delta > 0$ such that if $\mu(A) < \delta$, then $\int_A f d\mu < \xi/2$. Thus if we choose $\lambda \leq \delta^2(\xi)/4$, we obtain $|\alpha_n - \tilde{D}| \leq \xi$.

Substituting in (3.4), we obtain

$$
\begin{aligned}
Q(E_n) &\geq (1 - 2\sqrt{\lambda}) \exp\Big[- n \frac{\tilde{D} + \xi}{\tilde{\mu}(E_n)} \Big] \\
&\geq (1 - 2\sqrt{\lambda}) \exp\Big[- n \frac{\tilde{D} + \xi}{1 - 2\sqrt{\lambda}} \Big] \ . \qquad (3.5)
\end{aligned}
$$

On the other hand we have $E_n \subset \Gamma^k C_n \times \Gamma^k F_n \subset \Gamma^{2k}(C_n \times F_n) = \Gamma^{2k} G_n$. Thus if $\bar{z}_1^n \in \Gamma^{2k} G_n$, we can find a $z_1^n \in G_n$ such that $d_H(z_1^n, \bar{z}_1^n) \leq 2k_n$ and

$$
\begin{aligned}
Q(\bar{z}_1^n) &= \pi_V(\bar{z}_1) \prod_{i=2}^{n} V(\bar{z}_i | \bar{z}_{i-1}) \\
&\leq \rho^{-4k} \pi_V(z_1) \prod_{i=2}^{n} V(z_i | z_{i-1}) = Q(z_1^n) \rho^{-4k} \ ,
\end{aligned}
$$

where

$$
\rho \stackrel{\text{def}}{=} \min_{z_{-1}^0 \in \mathcal{Z}^2} V(z_0 | z_{-1}) \wedge \min_{z_{-1} \in \mathcal{Z}} \pi_V(z_{-1}) > 0 \ .
$$

This implies that

$$
Q(E_n) \leq \exp[n\nu(k_n/n)] Q(C_n \times F_n) \ , \qquad (3.6)
$$

where

$$\nu(u) \stackrel{\text{def}}{=} h(2u) + 2u \log \frac{|\mathcal{Z}|}{\rho^2} .$$

It is obvious that $\nu(u) \to 0$ as $u \to 0$. Since

$$k_n = \lceil n\sqrt{\lambda} \rceil \leq n\sqrt{\lambda} + 1 \qquad \text{and} \qquad h(u+v) \ \leq \ h(u) + h(v) \ ,$$

we have for sufficiently small λ and sufficiently large n,

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

Combining equations (3.5) and (3.6), we obtain

$$\begin{aligned} \beta_n(M,\epsilon) \ &= \ Q(\mathcal{A}_n) \ \geq \ Q(C_n \times F_n) \\ &\geq \ Q(E_n) \exp[-n\nu(k_n/n)] \\ &\geq \ (1 - 2\sqrt{\lambda}) \exp\left[- n(\frac{\tilde{D} + \xi}{1 - 2\sqrt{\lambda}} + \nu(\sqrt{\lambda}) + \nu(1/n))\right] \ . \end{aligned}$$

This yields

$$\begin{aligned} \theta(M,\epsilon) \ &\leq \ \limsup_n (-\frac{1}{n} \log \beta_n(M,\epsilon)) \\ &\leq \ \frac{\tilde{D} + \xi}{1 - 2\sqrt{\lambda}} + \nu(\sqrt{\lambda}) \\ &= \ \tilde{D} + \frac{2\tilde{D}\sqrt{\lambda} + \xi}{1 - 2\sqrt{\lambda}} + \nu(\sqrt{\lambda}) \ . \end{aligned}$$

The last expression can be made arbitrarily close to $\tilde{D}$ by choice of ξ and $\lambda \leq \delta^2(\xi)/4$. $\triangle$

The above two theorems yield the final result on the error exponent:

THEOREM 6.3. *For all $\epsilon \in (0,1)$ and $M \geq 2$, the error exponent for the hypothesis testing problem formulated in Section 6.2.(b) is given by*

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

where $\mathcal{L}$ is defined in equation (3.2). Moreover, for any null hypothesis W_2 there is a universal asymptotically optimal acceptance region that is independent of the alternative hypothesis V.

6.4. Extensions and concluding remarks

We obtained our results in the previous section under the following assumptions:

(1) the X-data are compressed into a fixed number of codewords and the Y-data are not compressed; and

(2) the process $\{(X_i, Y_i),\ i \in \mathbf{Z}\}$ is stationary ergodic (1st order) Markov under both hypotheses, such that $W = W_2$ is irreducible aperiodic under the null hypothesis and $V = V_2 > 0$ under the alternative.

These assumptions can be somewhat relaxed to yield similar results. In particular:

(a) If the X-data are compressed into a fixed number of codewords and the Y-data are compressed at any rate (with at least two codewords), then the error exponent remains unchanged.

(b) For all $k, l \in \{2, 3, \ldots\}$, let the observations be $(k-1)$th order Markov under the null hypothesis and $(l-1)$th order Markov under the alternative. If the transition matrix of the process under H_0 is denoted by $W = W_k$ (with $P = \pi_W \diamond W_k$ strongly mixing), and the transition matrix under H_1 is $V = V_l > 0$, then the error exponent is given by the infimum of the quantity

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

over all stationary distributions $\tilde{P}$ on $(\mathcal{X} \times \mathcal{Y})^{\mathbf{Z}}$ that satisfy for all $n \geq 0$ the conditions $\tilde{P}(x_{-n}^0) = P(x_{-n}^0)$ and $\tilde{P}(y_{-n}^0) = P(y_{-n}^0)$.

Finally we would like to emphasize that in the case of memoryless sources (viz. Chapter 2), the corresponding results were obtained under the wider assumption of *asymptotically zero rate*. In this chapter we have been unable to obtain results under this broader assumption because the version of the blowing-up lemma used for strongly mixing Markov sources was somewhat weaker than that used for memoryless sources.

APPENDIX A

Let θ' be defined by

$$\theta' \stackrel{\text{def}}{=} \{D(P_X||Q_X) \wedge D(\bar{P}_Y||Q_Y)\} \vee \{D(\bar{P}_X||Q_X) \wedge D(P_Y||Q_Y)\} .$$

For any $\Phi \subset \mathcal{P}(\mathcal{X})$, $\Psi \subset \mathcal{P}(\mathcal{Y})$, define $\alpha(\Phi, \Psi)$ by

$$\alpha(\Phi, \Psi) \stackrel{\text{def}}{=} \inf_{(\tilde{P}_X, \tilde{P}_Y) \in (\Phi \times \Psi) \cup (\Phi^c \times \Psi^c)} d(\tilde{P}_X, \tilde{P}_Y||Q) .$$

Referring to Theorem 3.4, we show that if

$$\theta^{(3)} \stackrel{\text{def}}{=} \sup_{\substack{\Phi, \Psi: \\ (P_X, P_Y) \in \Phi \times \Psi, \ (P_X, P_Y) \in \Phi^c \times \Psi^c}} \alpha(\Phi, \Psi) , \tag{A.1}$$

then $\theta^{(3)} = \theta'$.

(i) To show that $\theta^{(3)} \geq \theta'$, let $\Phi = \{P_X\}$ and $\Psi = \{\bar{P}_Y\}^c$. Then

$$\begin{aligned} \theta^{(3)} \geq \alpha(\Phi, \Psi) &= d(P_X, \{\bar{P}_Y\}^c||Q) \wedge d(\{P_X\}^c, \bar{P}_Y||Q) \\ &= D(P_X||Q_X) \wedge D(\bar{P}_Y||Q_Y) , \end{aligned} \tag{A.2}$$

where the last equality follows by continuity of divergence. Similarly, if $\Phi = \{\bar{P}_X\}^c$ and $\Psi = \{P_Y\}$, we have

$$\theta^{(3)} \geq D(P_Y||Q_Y) \wedge D(\bar{P}_X||Q_X) . \tag{A.3}$$

Combining (A.2) with (A.3) we obtain $\theta^{(3)} \geq \theta'$.

(ii) To show the reverse inequality $\theta^{(3)} \leq \theta'$, let

$$A = \text{cl}\Phi, \quad \bar{A} = \text{cl}\Phi^c, \quad B = \text{cl}\Psi, \quad \bar{B} = \text{cl}\Psi^c ,$$

where cl denotes closure under sup norm. Then by continuity of divergence,

$$\alpha(\Phi,\Psi) \;=\; \min_{(\tilde{P}_X,\tilde{P}_Y)\in(A\times B)\cup(\bar{A}\times\bar{B})} d(\tilde{P}_X,\tilde{P}_Y||Q)\;.$$

We must show that $\alpha(\Phi,\Psi)\le\theta'$ for every Φ and Ψ. This is trivially true if $(Q_X,Q_Y)\in(A\times B)\cup(\bar{A}\times\bar{B})$, in which case we have

$$\alpha(\Phi,\Psi) \;=\; d(Q_X,Q_Y||Q) \;=\; 0\;.$$

Hence we may assume that

$$(Q_X,Q_Y)\notin(A\times B)\cup(\bar{A}\times\bar{B})\;. \tag{A.4}$$

We provide an upper bound $\alpha(\Phi,\Psi)$ as follows. First we note that

$$(A\cap\bar{A})\times\mathcal{P}(\mathcal{Y}) \;\subset\; (A\times B)\cup(\bar{A}\times\bar{B})\;,$$

so that

$$\begin{aligned}\alpha(\Phi,\Psi) \;&\le\; \min_{(\tilde{P}_X,\tilde{P}_Y)\in(A\cap\bar{A})\times\mathcal{P}(\mathcal{Y})} d(\tilde{P}_X,\tilde{P}_Y||Q)\\ &=\; \min_{(\tilde{P}_X,\tilde{P}_Y):\;\tilde{P}_X\in A\cap\bar{A}} D(\tilde{P}_{XY}||Q_{XY})\;.\end{aligned}$$

Using the log-sum inequality, we can show that above minimum is equal to

$$\min_{\tilde{P}_X\in A\cap\bar{A}} D(\tilde{P}_X||Q_X)\;.$$

By symmetry we conclude that

$$\alpha(\Phi,\Psi)\le \min_{\tilde{P}_X\in A\cap\bar{A}} D(\tilde{P}_X||Q_X) \;\wedge\; \min_{\tilde{P}_Y\in B\cap\bar{B}} D(\tilde{P}_Y||Q_Y)\;. \tag{A.5}$$

Two cases may arise, according to whether Q_X lies in A or $\bar{A}$ (note that it cannot lie in $A\cap\bar{A}$ by (A.4)).

Case 1. $Q_X \in A$: Since $\bar{P}_X \in \bar{A}$, there exists $\lambda \in (0,1]$ such that

$$\hat{P}_X = \lambda \bar{P}_X + (1-\lambda) Q_X \in A \cap \bar{A} \; .$$

This yields

$$\begin{aligned} \min_{\tilde{P}_X \in A \cap \bar{A}} D(\tilde{P}_X || Q_X) \; &\leq \; D(\hat{P}_X || Q_X) \\ &\leq \; \lambda D(\bar{P}_X || Q_X) + (1-\lambda) D(Q_X || Q_X) \\ &\leq \; D(\bar{P}_X || Q_X) \; , \end{aligned}$$

where we used the convexity of divergence.

From (A.4), we also have that $Q_Y \in \bar{B}$. An analogous argument for $Q_Y \in \bar{B}$ and $P_Y \in B$ yields

$$\min_{\tilde{P}_Y \in B \cap \bar{B}} D(\tilde{P}_Y || Q_Y) \; \leq \; D(P_Y || Q_Y) \; .$$

From (A.5), we conclude that

$$\alpha(\Phi, \Psi) \leq D(\bar{P}_X || Q_X) \; \wedge \; D(P_Y || Q_Y) \; . \tag{A.6}$$

Case 2. $Q_X \in \bar{A}$: Again (A.4) implies that $Q_Y \in B$. As in Case 1 above, we obtain

$$\alpha(\Phi, \Psi) \leq D(P_X || Q_X) \; \wedge \; D(\bar{P}_Y || Q_Y) \; . \tag{A.7}$$

From (A.6) and (A.7) we conclude that $\alpha(\Phi, \Psi) \leq \theta'$, and hence also $\theta^{(3)} \leq \theta'$.

$\triangle$

APPENDIX B

Let $P_X \ll Q_X$. Define for any stochastic matrix $\Delta_{U|X}$ on $\mathcal{U} \times \mathcal{X}$ the following two distributions:

$$P_U(\cdot) \stackrel{\text{def}}{=} \sum_x \Delta_{U|X}(\cdot|x) P_X(x) = \sum_{x: Q_X(x)>0} \Delta_{U|X}(\cdot|x) P_X(x) \ ,$$

$$Q_U(\cdot) \stackrel{\text{def}}{=} \sum_{x: Q_X(x)>0} \Delta_{U|X}(\cdot|x) Q_X(x) \ .$$

Referring to the proof of Theorem 5.5, we show that if

$$\sigma^2 \stackrel{\text{def}}{=} \sup_{\Delta_{U|X}} \text{Var}_P \log \frac{P_U(U)}{Q_U(U)} \ ,$$

then $\sigma^2 < \infty$.

We define for any stochastic matrix $\Delta_{U|X}$ the functional

$$f(\Delta_{U|X}) \stackrel{\text{def}}{=} \sum_{u \in \mathcal{U}} P_U(u) \log^2 \frac{P_U(u)}{Q_U(u)} = \sum_{u: P_U(u)>0} P_U(u) \log^2 \frac{P_U(u)}{Q_U(u)} \ . \quad \text{(B.1)}$$

We have, for all $u \in \mathcal{U}$ with $P_U(u) > 0$,

$$P_U(u) \le \frac{P_U(u)}{Q_U(u)} \le \frac{\sum_{x: Q_X(x)>0} \Delta_{U|X}(u|x)}{\rho \sum_{x: Q_X(x)>0} \Delta_{U|X}(u|x)} = \frac{1}{\rho} \ ,$$

where $\rho \stackrel{\text{def}}{=} \min_{x: Q_X(x)>0} Q_X(x)$. Consequently,

$$\log P_U(u) \le \log \frac{P_U(u)}{Q_U(u)} \le \log \frac{1}{\rho}$$

and hence,

$$\log^2 \frac{P_U(u)}{Q_U(u)} \le \log^2 \frac{1}{\rho} \vee \log^2 P_U(u) \ .$$

Substituting in (B.1), we obtain

$$f(\Delta_{U|X}) \le \sum_{u: P_U(u)>0} \left(P_U(u) \log^2 \frac{1}{\rho}\right) \vee \left(P_U(u) \log^2 P_U(u)\right) \ .$$

Using the fact that $0 \le t \log^2 t \le \log^2 e^{2/e}$ for all $0 \le t \le 1$, it follows that

$$\begin{aligned} f(\Delta_{U|X}) &\le \sum_{u:P_U(u)>0} \log^2 \frac{1}{\rho} \vee \log^2 e^{2/e} \\ &\le |\mathcal{U}| \log^2 \left(\frac{1}{\rho} \vee e^{2/e}\right) . \end{aligned}$$

Hence

$$\begin{aligned} \sigma^2 &= \sup_{\Delta_{U|X}} \left\{ E_P \log^2 \frac{P_U(u)}{Q_U(u)} - \left(E_P \log \frac{P_U(u)}{Q_U(u)}\right)^2 \right\} \\ &\le \sup_{\Delta_{U|X}} f(\Delta_{U|X}) \le |\mathcal{U}| \log^2 \left(\frac{1}{\rho} \vee e^{2/e}\right) . \end{aligned}$$

$\triangle$

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc. London*, Series A-231, pp. 289-337, 1933.

[2] S. Kullback and R.A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79-86, 1951.

[3] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959 and New York: Dover, 1968.

[4] I. N. Sanov, "On the probability of large deviations of random variables," *Mat. Sbronik*, 42, pp. 11-44, 1957.

[5] W. Hoeffding, "Asymptotically optimal tests for multinominal distributions," *Ann. Math. Stat.*, vol. 36, pp. 369-400, 1965.

[6] I. Csiszár, "Sanov property, generalized I-projection, and a conditional limit theorem," *Ann. Prob.*, 12, pp. 768-793, 1984.

[7] I. Csiszár and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses," *Studia Sci. Math. Hungar.*, vol. 6, pp. 181-191, 1971.

[8] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Stat.*, vol. 23, pp. 493-507, 1952.

[9] T. Berger, "Decentralized estimation and decision theory," presented at the IEEE Seven Springs Workshop on Information Theory, Mt. Kisco, NY, September 1979.

[10] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 533-542, July 1986.

[11] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 759-772, Nov. 1987.

[12] T. S. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 2-14, Jan. 1989.

[13] S. I. Amari and T. S. Han, "Statistical inference under multiterminal rate restrictions: a differential geometric approach," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 217-227, Mar. 1989.

[14] I. Y. Hoballah and P. K. Varshney, "An information theoretic approach to the distributed detection problem," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 988-994, Sep. 1989.

[15] I. Y. Hoballah and P. K. Varshney, "Distributed Bayesian signal detection," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 995-1000, Sep. 1989.

[16] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 445-446, May 1986.

[17] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1982 and Budapest, Hungary: Akadémiai Kiadó, 1981.

[18] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *IEEE Trans. Aerospace and Electronic Sys.*, vol. AES-17, pp. 501-510, July 1981.

[19] J. J. Chao and C. C. Lee, "A distributed detection scheme based on soft local decisions," *the 24th Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois*, Oct. 1-3, 1986.

[20] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerospace and Electronic Sys.*, vol. AES-22, pp. 98-101, Jan. 1986.

[21] S. C. Thomopoulos, R. Viswanathan, and D. C. Bougoulias, "Optimal decision fusion in multiple sensor systems," *IEEE Trans. Aerospace and Electronic Sys.*, vol. AES-23, pp. 644-653, Sep. 1987.

[22] J. N. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 440-446, May 1985.

[23] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals, and Systems*, vol. 1,2, pp. 167-182, 1988.

[24] A. R. Reibman and L. W. Nolte, "Optimal detection and performance of distributed sensor systems," *IEEE Trans. Aerospace and Electronic Sys.*, vol. AES-23, pp. 24-30, Jan. 1987.

[25] E. Geraniotis, "Robust distributed discrete-time block and sequential detection," *in proc. 1987 Conf. Inform Sci. Syst.*, Johns Hopkins Univ., pp. 354-360, Mar. 1987.

[26] H. R. Hashemi and I. B. Rhodes, "Decentralized sequential detection," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 509-520, May 1989.

[27] E. Geraniotis and Y. A. Chau, "Distributed detection of weak signals from multiple sensors with correlated observations," *in proc. of the 27th Conf. on Detection and Control*, Austin, Texas, pp. 2501-2506, Dec. 1988.

[28] E. Geraniotis and Y. A. Chau, "Robust data fusion for multisensor detection systems," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 1265-1279, Nov. 1990.

[29] D. Teneketzis and P. Varaiya, "The decentralized quickest decision problem," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 641-644, July 1984.

[30] J. J. Chao and C. C. Lee, "Optimal Local Decision Space Partitioning for Distributed Detection," *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-25, pp. 536-543, July 1989.

[31] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 198-211, Mar. 1988.

[32] I. Y. Hoballah and P. K. Varshney, "Neyman-Pearson detection with distributed sensors," *In Proceedings of the 25th IEEE Conference on Decision and Control,* Athens, Greece, Dec. 1986, pp. 237-241.

[33] G. Polychronopoulos and J. N. Tsitsiklis, "Explicit solutions for some simple decentralized detection problems," *IEEE Trans. Aerospace and Electronic Sys.*, vol. AES-26, pp. 282-292, Mar. 1990.

[34] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press, 1967.

[35] 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.

[36] 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.

[37] 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.

[38] P. C. Shields, "Stationary Coding of Processes," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 283-291, 1979.

[39] 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.

[40] N. A. Friedman and D. S. Ornstein, "On Isomorphism of weak Bernoulli transformations," *Advances in Math.*, vol. 5, pp. 365-394, 1971.

[41] 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.

[42] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196-219, 1953.

[43] S. C. Moy, "Generalizations of Shannon-McMillan theorem," *Pacific Jour. Math*, vol. 11, pp. 705-714, 1961.

[44] H. L. Royden, *Real Analysis,* 2nd ed. New York: Macmillan Publishing Co., Inc., 1968.

[45] A. N. Shiryayev, *Probability.* New York: Springer-Verlag, 1984.