The Method of Types

Imre Csiszár, *Fellow, IEEE*

(Invited Paper)

***Abstract*— The method of types is one of the key technical tools in Shannon Theory, and this tool is valuable also in other fields. In this paper, some key applications will be presented in sufficient detail enabling an interested nonspecialist to gain a working knowledge of the method, and a wide selection of further applications will be surveyed. These range from hypothesis testing and large deviations theory through error exponents for discrete memoryless channels and capacity of arbitrarily varying channels to multiuser problems. While the method of types is suitable primarily for discrete memoryless models, its extensions to certain models with memory will also be discussed.**

***Index Terms*—Arbitrarily varying channels, choice of decoder, counting approach, error exponents, extended type concepts, hypothesis testing, large deviations, multiuser problems, universal coding.**

I. Introduction

ONE of Shannon's key discoveries was that—for quite general source models—the negative logarithm of the probability of a typical long sequence divided by the number of symbols is close to the source entropy H; the total probability of all n-length sequences not having this property is arbitrarily small if n is large. Thus "it is possible for most purposes to treat long sequences as though there were just 2^{Hn} of them, each with a probability 2^{-Hn}" [75, p. 24]. Shannon demonstrated the power of this idea also in the context of channels. It should be noted that Shannon [75] used the term "typical sequence" in an intuitive rather than technical sense. Formal definitions of typicality, introduced later, need not concern us here.

At the first stage of development of information theory, the main theoretical issue was to find the best rates of source or channel block codes that, assuming a known probabilistic model, guarantee arbitrarily small probability of error (or tolerable average distortion) when the blocklength n is sufficiently large. For this purpose, covered by the previous quotation from Shannon [75], typical sequences served as a very efficient and intuitive tool, as demonstrated by the book of Wolfowitz [81].

The limitations of this tool became apparent when interest shifted towards the speed of convergence to zero of the error probability as $n \to \infty$. Major achievements of the 1960's were, in the context of discrete memoryless channels (DMC's), the "random coding" upper bound and the "sphere packing" lower bound to the error probability of the best code of a given rate less than capacity (Fano [44], Gallager [46], Shannon, Gallager, and Berlekamp [74]). These bounds exponentially coincide for rates above a "critical rate" and provide the exact error exponent of a DMC for such rates. These results could not be obtained via typical sequences, and their first proofs used analytic techniques that gave little insight.

It turned out in the 1970's that a simple refinement of the typical sequence approach is effective—at least in the discrete memoryless context—also for error exponents, as well as for situations where the probabilistic model is partially unknown. The idea of this refinement, known as the method of types, is to partition the n-length sequences into classes according to type (empirical distribution). Then the error event of interest is partitioned into its intersections with the type classes, and the error probability is obtained by summing the probabilities of these intersections. The first key fact is that the number of type classes grows subexponentially with n. This implies that the error probability has the same exponential asymptotics as the largest one among the probabilities of the above intersections. The second key fact is that sequences of the same type are equiprobable under a memoryless probabilistic model. Hence to bound the probabilities of intersections as above it suffices to bound their cardinalities, which is often quite easy. This informal description assumes models involving one set of sequences (source coding or hypothesis testing); if two or more sets of sequences are involved (as in channel coding), joint types have to be considered.

In this paper, we will illustrate the working and the power of the method of types via a sample of examples that the author considers typical and both technically and historically interesting. The simple technical background, including convenient notation, will be introduced in Section II. The first key applications, viz. universally attainable exponential error bounds for hypothesis testing and channel block-coding, will be treated in Sections III and IV, complete with proofs. The universally attainable error exponent for source block-coding arises as a special case of the hypothesis testing result. A basic result of large deviations theory is also included in Section III. Section V is devoted to the arbitrarily varying channel (AVC) capacity problem. Here proofs could not be given in full, but the key steps are reproduced in detail showing how the results were actually obtained and how naturally the method of types suggested a good decoder. Other typical applications are reviewed in Section VI, including rate-distortion theory, source-channel error exponents, and

Manuscript received December 16, 1997; revised April 20, 1998. This work was supported by the Hungarian National Foundation for Scientific Research under Grant T016386.

The author is with the Mathematical Institute of the Hungarian Academy of Sciences, H1364 Budapest, P.O. Box 127, Hungary.

Publisher Item Identifier S 0018-9448(98)05285-7.

multiuser problems. Although the method of types is tailored to discrete memoryless models, there exist extensions of the type concept suitable for certain models with memory. These will be discussed in Section VII.

The selection of problems and results treated in this paper has been inevitably subjective. To survey all applications of the method of types would have required a paper the size of a book. In particular, several important applications in Combinatorics are not covered, in this respect the reader should consult the paper of Körner and Orlitsky in this issue.

While historical aspects were taken seriously, and a rather large list of references has been included, no attempts were made to give a detailed account of the history of the method of types. About its origins let us just make the following brief comments.

The ingredients have been around for a long time. In probability theory, they appear in the basic papers of Sanov [72] and Hoeffding [53] on large deviations, cf. Section III below. A similar counting approach had been used in statistical physics even earlier, dating back to Boltzmann [18]. A remarkable example is the paper of Schrödinger [73] that predates modern large deviations theory but remained unknown outside the physics community until recently. Information theorists have also used ideas now considered pertinent to the method of types. Fano's [44] approach to the DMC error exponent problem was based on "constant composition codes," and Berger's [13] extension of the rate-distortion theorem to sources with partially known and variable statistics relied upon his key lemma about covering a type class. Later, in the 1970's, several authors made substantial use of the concept now called joint type, including Blahut [16], Dobrushin and Stambler [40], and Goppa [48].

While the ideas of the method of types were already around in the 1970's, this author believes that his research group is fairly credited for developing them to a general method, indeed, to a basic tool of the information theory of discrete memoryless systems. The key coworkers were János Körner and Katalin Marton. A systematic development appears in the book of Csiszár and Körner [30]. Were that book written now, both authors would prefer to rely even more extensively on types, rather than typical sequences. Indeed, while "merging nearby types, i.e., the formalism of typical sequences has the advantage of shortening computations" [30, p. 38], that advantage is relatively minor in the discrete memoryless context. On the other hand, the less delicate "typical sequence" approach is more robust, it can be extended also to those models with memory or continuous alphabets for which the type idea apparently fails.

II. Technical Background

The technical background of the method of types is very simple. In the author's information theory classes, the lemmas below are part of the introductory material.

$\mathcal{X}, \mathcal{Y}, \cdots$ will denote finite sets, unless stated otherwise; the size of $\mathcal{X}$ is denoted by $|\mathcal{X}|$. The set of all probability distributions (PD's) on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. For PD's P and Q, $H(P)$ denotes entropy and $D(P\|Q)$ denotes information divergence, i.e.,

$$H(P) = -\sum_{a\in\mathcal{X}} P(a)\log P(a)$$

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(a)\log\frac{P(a)}{Q(a)}$$

with the standard conventions that $0\log 0 = 0\log 0/0 = 0$, $p\log(p/0) = \infty$ if $p > 0$. Here and in the sequel the base of log and of exp is arbitrary but the same; the usual choices are 2 or e.

The type of a sequence $\boldsymbol{x} = x_1\cdots x_n \in \mathcal{X}^n$ and the joint type of $\boldsymbol{x}$ and $\boldsymbol{y} = y_1\cdots y_n \in \mathcal{Y}^n$ are the PD's $P_{\boldsymbol{x}} \in \mathcal{P}(\mathcal{X})$ and $P_{\boldsymbol{xy}} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ defined by letting $P_{\boldsymbol{x}}(a)$ and $P_{\boldsymbol{xy}}(a, b)$ be the relative frequency of a among $x_1, \cdots, x_n$ and of (a, b) among $(x_1, y_1), \cdots, (x_n, y_n)$, respectively, for all $a \in \mathcal{X}$, $b \in \mathcal{Y}$. Joint types of several n-length sequences are defined similarly. The subset of $\mathcal{P}(\mathcal{X})$ consisting of the possible types of sequences $\boldsymbol{x} \in \mathcal{X}^n$ is denoted by $\mathcal{P}_n(\mathcal{X})$.

Lemma II.1:

$$|\mathcal{P}_n(\mathcal{X})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}.$$

Proof: Elementary combinatorics.

The probability that n independent drawings from a PD $Q \in \mathcal{P}(\mathcal{X})$ give $\boldsymbol{x} \in \mathcal{X}^n$, is denoted by $Q^n(\boldsymbol{x})$. Similarly, the probability of receiving $\boldsymbol{y} \in \mathcal{Y}^n$ when $\boldsymbol{x} \in \mathcal{X}^n$ is sent over a DMC with matrix W, is denoted by $W^n(\boldsymbol{y}|\boldsymbol{x})$. Clearly, if $\boldsymbol{x} \in \mathcal{X}^n$ have type P and $(\boldsymbol{x}, \boldsymbol{y})$ have joint type $\tilde{P}$

$$\begin{aligned} Q^n(\boldsymbol{x}) &= \prod_{a\in\mathcal{X}} Q(a)^{nP(a)} \\ &= \exp\{-n[H(P) + D(P\|Q)]\} \qquad \text{(II.1)} \\ W^n(\boldsymbol{y}|\boldsymbol{x}) &= \prod_{a\in\mathcal{X},\, b\in\mathcal{Y}} W(b|a)^{n\tilde{P}(a,\,b)} \\ &= \exp\{-n[H(\tilde{P}) - H(P) + D(\tilde{P}\|W)]. \qquad \text{(II.2)} \end{aligned}$$

Here $D(\tilde{P}\|W)$ is defined for any $\tilde{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ by

$$\begin{aligned} D(\tilde{P}\|W) &= D(\tilde{P}\|P\times W) \\ (P\times W)(a, b) &= P(a)W(b|a) \qquad \text{(II.3)} \end{aligned}$$

where P denotes the $\mathcal{X}$-marginal of $\tilde{P}$.

We will write $P \ll Q$, respectively $\tilde{P} \ll W$, to denote that P or $\tilde{P}$ is 0 for each $a \in \mathcal{X}$ or $(a, b) \in \mathcal{X}\times\mathcal{Y}$ with $Q(a) = 0$ or $W(b|a) = 0$ respectively. The divergences in (II.1) and (II.2) are finite iff $P \ll Q$, respectively, $\tilde{P} \ll W$.

For $P \in \mathcal{P}_n(\mathcal{X})$, the type class $\{\boldsymbol{x} \in \mathcal{X}^n, P_{\boldsymbol{x}} = P\}$ will be denoted by $\mathcal{T}_P^n$. Similarly, for $\tilde{P} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y})$ we write $\mathcal{T}_{\tilde{P}}^n = \{(\boldsymbol{x}, \boldsymbol{y})\colon \boldsymbol{x} \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{Y}^n, P_{\boldsymbol{xy}} = \tilde{P}\}$.

Lemma II.2: For any type $P \in \mathcal{P}_n(\mathcal{X})$

$$|\mathcal{P}_n(\mathcal{X})|^{-1}\exp\{nH(P)\} \le |\mathcal{T}_P^n| \le \exp\{nH(P)\} \qquad \text{(II.4)}$$

and for any PD $Q \in \mathcal{P}(\mathcal{X})$

$$\begin{aligned} |\mathcal{P}_n(\mathcal{X})|^{-1}\exp\{-nD(P\|Q)\} &\le Q^n(\mathcal{T}_P^n) \\ &\le \exp\{-nD(P\|Q)\}. \qquad \text{(II.5)} \end{aligned}$$

Proof: Equation (II.1) with $Q = P$ gives

$$P^n(\mathcal{T}_P^n) = |\mathcal{T}_P^n| \exp\{-nH(P)\}.$$

Hence (II.4) follows because

$$\begin{aligned} 1 \geq P^n(\mathcal{T}_P^n) &= \max_{P' \in \mathcal{P}_n(\mathcal{X})} P^n(\mathcal{T}_{P'}^n) \\ &\geq |\mathcal{P}_n(\mathcal{X})|^{-1} \sum_{P' \in \mathcal{P}_n(\mathcal{X})} P^n(\mathcal{T}_{P'}^n) \\ &= |\mathcal{P}_n(\mathcal{X})|^{-1} \end{aligned}$$

where the first equality can be checked by simple algebra. Clearly, (II.1) and (II.4) imply (II.5).

Remark: The bounds (II.4) and (II.5) could be sharpened via Stirling approximation to factorials, but that sharpening is seldom needed.

In the sequel, we will use the convenient notations $\lesssim$ and $\approx$ for inequality and equality up to a polynomial factor, i.e., $f(n) \lesssim g(n)$ means that $f(n) \leq p(n)g(n)$ for all n, where $p(n)$ is some polynomial of n, and $f(n) \approx g(n)$ means that both $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. When $f(n)$ and $g(n)$ depend not only on n but on other variables as well, it is understood that the polynomial $p(n)$ can be chosen independently of those. With this notation, by Lemmas II.1 and II.2, we can write

$$\begin{aligned} |\mathcal{T}_P^n| &\approx \exp\{nH(P)\} \\ Q^n(\mathcal{T}_P^n) &\approx \exp\{-nD(P\|Q)\}. \end{aligned} \tag{II.6}$$

Random variables (RV's) with values in $\mathcal{X}$, $\mathcal{Y}$, etc., will be denoted by $X, Y, \cdots$ (often with indices). Distributions, respectively, joint distributions of RV's are denoted by P_X, P_{XY}, etc. It is often convenient to represent types, particularly joint types, as (joint) distributions of dummy RV's. For dummy RV's with $P_X = P \in \mathcal{P}_n(\mathcal{X})$ or $P_{XY} = \tilde{P} \in \mathcal{P}_n(\mathcal{X}\times)$, etc., we will write $\mathcal{T}_X^n$, $\mathcal{T}_{XY}^n$, etc., instead of $\mathcal{T}_P^n$, $\mathcal{T}_{\tilde{P}}^n$, etc. Also, we will use notations like $\mathcal{T}_{Y|X}^n(\boldsymbol{x}) = \{\boldsymbol{y}\colon (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{T}_{XY}^n\}$.

Lemma II.3: For X, Y representing a joint type, i.e., $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$, and any $\boldsymbol{x} \in \mathcal{T}_X^n$ and channel W

$$\begin{aligned} |\mathcal{T}_{Y|X}^n(\boldsymbol{x})| &\approx \exp\{nH(Y|X)\} \\ W^n(\mathcal{T}_{Y|X}^n(\boldsymbol{x})|\boldsymbol{x}) &\approx \exp\{-nD(P_{XY}\|W)\}. \end{aligned} \tag{II.7}$$

Proof: As $|\mathcal{T}_{Y|X}^n(\boldsymbol{x})|$ is constant for $\boldsymbol{x} \in \mathcal{T}_X^n$, it equals $|\mathcal{T}_{XY}^n|/|\mathcal{T}_X^n|$. Thus the first assertion follows from (II.4), since $H(P_{XY}) - H(P_X) = H(Y|X)$. The second assertion follows from the first one and (II.2).

The representation of types by dummy RV's suggests the introduction of "information measures" for (nonrandom) sequences. Namely, for $\boldsymbol{x} \in \mathcal{X}^n$, $\boldsymbol{y} \in \mathcal{Y}^n$, we define the (nonprobabilistic or empirical) entropy $H(\boldsymbol{x})$, conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$, and mutual information $I(\boldsymbol{x} \wedge \boldsymbol{y})$ as $H(X)$, $H(Y|X)$, $I(X \wedge Y)$ for dummy RV's X, Y whose joint distribution P_{XY} equals the joint type $P_{\boldsymbol{xy}}$. Of course, nonprobabilistic conditional mutual information like $I(\boldsymbol{x} \wedge \boldsymbol{y}|\boldsymbol{z})$ is defined similarly. Notice that on account of (II.1), for any $\boldsymbol{x} \in \mathcal{X}^n$ the probability $Q^n(\boldsymbol{x})$ is maximum if $Q = P_{\boldsymbol{x}}$. Hence the nonprobabilistic entropy $H(\boldsymbol{x}) = H(P_{\boldsymbol{x}})$ is actually the maximum-likehood estimate of the entropy $H(Q)$ based upon the observed sequence $\boldsymbol{x}$.

III. Source Block Coding, Hypothesis Testing, Large Deviation Theory

The working of the method of types is particularly simple in problems that involve only one set of sequences, such as source coding and hypothesis testing. Theorems III.1 and III.2 below establish the existence of source block codes and tests of (in general, composite) hypotheses, universally optimal in the sense of error exponents. Theorem III.1 appeared as a first illustration of the method of types in Csiszár and Körner [30, p. 37], cf. also Longo and Sgarro [63]. Formally, as pointed out below, it is a special case of Theorem III.2. The latter is effectively due to Hoeffding [53].

Theorem III.1: Given $0 < R < \log|\mathcal{X}|$, the sets $A_n = \{\boldsymbol{x}\colon H(\boldsymbol{x}) \leq R\} \subset \mathcal{X}^n$ satisfy

$$\frac{1}{n} \log |A_n| \to R \tag{III.1}$$

and for every PD $Q \in \mathcal{P}(\mathcal{X})$

$$\frac{1}{n} \log Q^n(A_n^c) \to -e_Q(R) \tag{III.2}$$

where

$$\begin{aligned} e_Q(R) &= \min_{P\colon H(P) \geq R} D(P\|Q) \\ e_Q(R) &> 0, \qquad \text{if } H(Q) < R. \end{aligned} \tag{III.3}$$

Moreover, for any sequence of sets $\tilde{A}_n \subset \mathcal{X}^n$ satisfying (III.1), we have for all $Q \in \mathcal{P}(\mathcal{X})$

$$\liminf_{n\to\infty} \frac{1}{n} \log Q^n(\tilde{A}_n^c) \geq -e_Q(R). \tag{III.4}$$

Interpretation: Encoding n-length sequences $\boldsymbol{x} \in \mathcal{X}^n$ by assigning distinct codewords to sequences of empirical entropy $\leq R$, this code is universally optimal among those of (asymptotic) rate R, for the class of memoryless sources: for any source distribution Q of entropy less than R, the error probability goes to 0 exponentially, with exponent that could not be improved even if the distribution Q were known.

For a set $\Pi \subset \mathcal{P}(\mathcal{X})$ of PD's, and $Q \in \mathcal{P}(\mathcal{X})$, write

$$D(\Pi\|Q) = \inf_{P \in \Pi} D(P\|Q). \tag{III.5}$$

Further, for $P \in \mathcal{P}(\mathcal{X})$ and $\alpha > 0$, denote by $B(P, \alpha)$ the "divergence ball with center P and radius α," and for $\Pi \subset \mathcal{P}(\mathcal{X})$ denote by $B(\Pi, \alpha)$ the "divergence α-neighborhood" of Π

$$\begin{aligned} B(P, \alpha) &= \{P'\colon D(P'\|P) < \alpha\} \\ B(\Pi, \alpha) &= \bigcup_{P \in \Pi} B(P, \alpha). \end{aligned} \tag{III.6}$$

Theorem III.2: Given any set $\Pi \subset \mathcal{P}(\mathcal{X})$ of PD's and $\alpha \geq 0$, let $A_n \subset \mathcal{X}^n$ be the set of those $\boldsymbol{x} \in \mathcal{X}^n$ whose type $P_{\boldsymbol{x}}$ is in the complement of $B(\Pi, \alpha)$ if $\alpha > 0$, respectively, in the complement of $B(\Pi, \alpha_n)$ if $\alpha = 0$ where $\alpha_n \to 0$, $\alpha_n n / \log n \to \infty$. Then

$$\sup_{P \in \Pi} P^n(A_n) \to 0, \quad \limsup_{n \to \infty} \frac{1}{n} \log \sup_{P \in \Pi} P^n(A_n) \leq -\alpha \tag{III.7}$$

and for every $Q \in \mathcal{P}(\mathcal{X})$

$$\lim_{n \to \infty} \frac{1}{n} \log Q^n(A_n^c) = \begin{cases} -D(B(\Pi, \alpha) \| Q), & \text{if } \alpha > 0 \\ -D(\overline{\Pi} \| Q), & \text{if } \alpha = 0 \end{cases} \tag{III.8}$$

where $\overline{\Pi}$ denotes the closure of Π.

Moreover, for arbitrary P and Q in $\mathcal{P}(\mathcal{X})$ and any sequence of sets $\tilde{A}_n \subset \mathcal{X}^n$

$$\limsup_{n \to \infty} \frac{1}{n} \log P^n(\tilde{A}_n) \leq -\alpha < 0$$

implies

$$\liminf_{n \to \infty} \frac{1}{n} \log Q^n(\tilde{A}_n^c) \geq -D(B(P, \alpha) \| Q) \tag{III.9}$$

and

$$\limsup_{n \to \infty} P^n(\tilde{A}_n) \leq 1$$

implies

$$\liminf_{n \to \infty} \frac{1}{n} \log Q^n(\tilde{A}_n^c) \geq -D(P \| Q). \tag{III.10}$$

Interpretation: For testing the null-hypothesis that the true distribution belongs to Π, with type 1 error probability required to decrease with a given exponential rate or just go to zero, a universally rate optimal test is the one with critical region A_n (i.e., the test rejects the null-hypothesis iff $\boldsymbol{x} \in A_n$). Indeed, by (III.9), (III.10), no tests meeting the type 1 error requirement (III.7), even if designed against a particular alternative Q, can have type 2 error probability decreasing with a better exponential rate than that in (III.8) (when $\alpha > 0$, this follows simply because

$$D(B(\Pi, \alpha) \| Q) = \inf_{P \in \Pi} D(B(P, \alpha) \| Q)$$

by (III.5) and (III.6); in the case $\alpha = 0$, one needs the observation that

$$\sup_{P \in \Pi} P^n(A) = \sup_{P \in \overline{\Pi}} P^n(A)$$

for every $A \subset \mathcal{X}^n$). In particular, for Q such that the exponent in (III.8) is 0, the type 2 error probability cannot exponentially decrease at all. The notion that the null-hypothesis is rejected whenever the type of the observed sample $\boldsymbol{x}$ is outside a "divergence neighborhood" of Π is intuitive enough. In addition, notice that by Lemma II.1 the rejection criterion $P_{\boldsymbol{x}} \notin B(\Pi, \alpha)$ is equivalent to

$$\frac{\sup_{P \in \Pi} P^n(\boldsymbol{x})}{\sup_{P \in \mathcal{P}(\mathcal{X})} P^n(\boldsymbol{x})} = \exp\left\{-n \inf_{P \in \Pi} D(P_{\boldsymbol{x}} \| P)\right\} \leq \exp(-n\alpha).$$

Hence the above universally rate optimal tests are what statisticians call likehood ratio tests.

Remarks:

i) One sees from (III.9) that in the case $\alpha > 0$, the type 2 error probability exponent (III.8) could not be improved even if (III.7) were replaced by the weaker requirement that

$$\limsup_{n \to \infty} \frac{1}{n} \log P^n(A_n) \leq -\alpha, \qquad \text{for each } P \in \Pi.$$

ii) In the case $\alpha = 0$, the requirement (III.7) could be relaxed to

$$\limsup_{n \to \infty} P^n(A_n) < 1, \qquad \text{for each } P \in \Pi$$

provided that $D(\overline{\Pi} \| Q) = D(\Pi \| Q)$; the latter always holds if $P \ll Q$ for each $P \in \Pi$ but not necessarily otherwise.

A particular case of this result is known as Stein's lemma, cf. Chernoff [20]: if a simple null-hypothesis P is to be tested against a simple alternative Q, with an arbitrarily fixed upper bound on the type 1 error probability, the type 2 error probability can be made to decrease with exponential rate $D(P \| Q)$ but not better.

iii) Theorem III.2 contains Theorem III.1 as the special case $\Pi = \{P_0\}$, $\alpha = \log |\mathcal{X}| - R$, where P_0 is the uniform distribution on $\mathcal{X}$.

Proof of Theorem III.2: Suppose first that $\alpha > 0$. Then A_n is the union of type classes $\mathcal{T}_{P'}^n$ for types P' not belonging to $B(P, \alpha)$ whenever $P \in \Pi$. By the definition (III.6) of $B(P, \alpha)$, for such P' we have $D(P' \| P) \geq \alpha$ and hence, by Lemma II.2, $P^n(\mathcal{T}_{P'}^n) \leq \exp(-n\alpha)$ for each $P \in \Pi$. This gives by Lemma II.1

$$P^n(A_n) \lesssim \exp(-n\alpha), \qquad \text{for } P \in \Pi \tag{III.11}$$

establishing (III.7) (the notation $\lesssim$ has been defined in Section II). Further, A_n^c is the union of type classes $\mathcal{T}_{P'}^n$ for types P' that belong to $B(\Pi, \alpha)$, thus satisfy $D(P' \| Q) \geq D(B(\Pi, \alpha) \| Q)$. Hence we get as above

$$Q^n(A_n^c) \lesssim \exp\{-nD(B(\Pi, \alpha) \| Q)\} \tag{III.12}$$

and this gives

$$\limsup_{n \to \infty} \frac{1}{n} \log Q_n(A_n^c) \leq -D(B(\Pi, \alpha) \| Q). \tag{III.13}$$

To complete the proof in the case $\alpha > 0$, it suffices to prove (III.9). Given any $0 < \alpha' < \alpha$, the assumption in (III.9) implies for n sufficiently large that

$$|\mathcal{T}_{P'}^n \cap \tilde{A}_n^c| \geq \tfrac{1}{2} |\mathcal{T}_{P'}^n|, \qquad \text{for all } P' \in B(P, \alpha') \cap \mathcal{P}_n(\mathcal{X}). \tag{III.14}$$

Indeed, else $|\mathcal{T}_{P'}^n \cap \tilde{A}_n| > \frac{1}{2} |\mathcal{T}_{P'}^n|$ would hold for some $P' \in B(P, \alpha') \cap \mathcal{P}_n(\mathcal{X})$. Since sequences in the same type class are equiprobable, the latter would imply by Lemma II.2 that

$$\begin{aligned} P^n(\tilde{A}_n) &\geq P^n(\mathcal{T}_{P'}^n \cap \tilde{A}_n) \\ &\geq \tfrac{1}{2} P^n(\mathcal{T}_{P'}^n) \approx \exp\{-nD(P' \| P)\} \gtrsim \exp(-n\alpha') \end{aligned}$$

contradicting the assumption in (III.9).

Given any $\varepsilon > 0$, take $\alpha' < \alpha$ such that

$$D(B(P, \alpha')\|Q) \le D(B(P, \alpha)\|Q) + \varepsilon$$

and take $P' \in B(P, \alpha') \cap \mathcal{P}_n(\mathcal{X})$ such that

$$D(P'\|Q) < D(B(P, \alpha')\|Q) + \varepsilon$$

(possible for large n). Then by (III.14) and (II.6)

$$\begin{aligned} Q^n(A_n^c) &\ge Q^n(\mathcal{T}_{P'}^n \cap A_n^c) \ge \tfrac{1}{2} Q^n(\mathcal{T}_{P'}^n) \\ &\approx \exp\{-nD(P'\|Q)\} \\ &\ge \exp\{-n(D(B(P, \alpha)\|Q) - 2\varepsilon)\}. \end{aligned} \quad \text{(III.15)}$$

As $\varepsilon > 0$ was arbitrary, (III.15) establishes (III.9).

In the remaining case $\alpha = 0$, (III.11) will hold with α replaced by α_n; using the assumption $\alpha_n n / \log n \to \infty$, this yields (III.7). Also (III.12) will hold with α replaced by α_n. It is easy to see that $\alpha_n \to 0$ implies

$$D(B(\Pi, \alpha_n)\|Q) \to D(\overline{\Pi}\|Q)$$

hence we get

$$\limsup_{n\to\infty} \frac{1}{n} \log Q^n(A_n^c) \le -D(\overline{\Pi}\|Q). \quad \text{(III.16)}$$

To complete the proof, it suffices to prove (III.10). Now, $P^n(\tilde{A}_n) < 1 - \varepsilon$ implies, for large n, that

$$|\mathcal{T}_{P'}^n \cap \tilde{A}_n^c| \ge \frac{\varepsilon}{2} |\mathcal{T}_{P'}^n|, \qquad \text{for some } P' \in B(P, \alpha_n) \cap \mathcal{P}_n(\mathcal{X}). \quad \text{(III.17)}$$

Indeed, else $|\mathcal{T}_{P'}^n \cap \tilde{A}_n| > (1 - (\varepsilon/2))|\mathcal{T}_{P'}^n|$ would hold for all $P' \in B(P, \alpha_n) \cap \mathcal{P}_n(\mathcal{X})$, implying

$$\begin{aligned} P^n(\tilde{A}_n) &\ge \sum_{P' \in B(P, \alpha_n) \cap \mathcal{P}_n(\mathcal{X})} P^n(\tilde{A}_n \cap \mathcal{T}_{P'}^n) \\ &\ge \left(1 - \frac{\varepsilon}{2}\right) P^n \left(\bigcup_{P' \in B(P, \alpha_n) \cap \mathcal{P}_n(\mathcal{X})} \mathcal{T}_{P'}^n \right). \end{aligned}$$

For large n, this contradicts $P^n(\tilde{A}_n) < 1 - \varepsilon$, since $\alpha_n n / \log n \to \infty$ implies by Lemmas II.1 and II.2 that the P^n-probability of the union of type classes with $P' \notin B(P, \alpha_n)$ goes to 0 as $n \to \infty$.

Pick $P_n' \in B(P, \alpha_n) \cap \mathcal{P}_n(\mathcal{X})$ satisfying (III.17), then

$$\begin{aligned} Q^n(A_n^c) &\ge Q^n(\mathcal{T}_{P_n'}^n \cap A_n^c) \ge \frac{\varepsilon}{2} Q^n(\mathcal{T}_{P_n'}^n) \\ &\approx \exp\{-nD(P_n'\|Q)\}. \end{aligned} \quad \text{(III.18)}$$

Here $D(P_n'\|P) < \alpha_n \to 0$ by assumption, and this implies $D(P_n'\|Q) \to D(P\|Q)$. Thus (III.18) gives (III.10).

Large Deviations, Gibbs' Conditioning Principle

Large deviation theory is concerned with "small probability events," typically of probability going to zero exponentially. An important example of such events is that the type of an independent and identically distributed (i.i.d.) random sequence $X^n = (X_1, \cdots, X_n)$ belongs to a given set $\Pi \subset \mathcal{P}(\mathcal{X})$ of PD's on $\mathcal{X}$ that does not contain the distribution Q of the RV's X_i. In this context, the type of X^n is called the empirical distribution $\tilde{P}_n$. Thus

$$\Pr\{\tilde{P}_n \in \Pi\} = Q^n(\{\boldsymbol{x}\colon P_{\boldsymbol{x}} \in \Pi\}). \quad \text{(III.19)}$$

Theorem III.3: For any $\Pi \subset \mathcal{P}(\mathcal{X})$

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr\{\tilde{P}_n \in \Pi\} \le -D(\Pi\|Q) \quad \text{(III.20)}$$

and if Π has the property that

$$\lim_{n\to\infty} D(\Pi \cap \mathcal{P}_n(\mathcal{X})\|Q) = D(\Pi\|Q) \quad \text{(III.21)}$$

then

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\tilde{P}_n \in \Pi\} = -D(\Pi\|Q). \quad \text{(III.22)}$$

Corollary: If Π satisfies (III.21) and a unique $P^* \in \overline{\Pi}$ satisfies $D(P^*\|Q) = D(\Pi\|Q)$ then for any fixed k, the conditional joint distribution of $X_1, \cdots, X_k$ on the condition $\tilde{P}_n \in \Pi$ approaches P^{*k} as $n \to \infty$.

Remarks:

i) The condition (III.21) is trivially satisfied if Π is an open subset of $\mathcal{P}(\mathcal{X})$ or, more generally, if each $P \in \Pi$ with $P \ll Q$ is contained in the closure of the set of those $\tilde{P} \ll Q$ for which all $P' \ll \tilde{P}$ sufficiently close to $\tilde{P}$ belong to Π. In particular, (III.8) for $\alpha > 0$ is an instance of (III.22). Hoeffding [53] considered sets of PD's Π such that (III.21) holds with rate of convergence $O(\log n/n)$. For such Π, (III.22) can be sharpened to

$$\Pr\{\tilde{P}_n \in \Pi\} \approx \exp\{-nD(\Pi\|Q)\}. \quad \text{(III.23)}$$

ii) The empirical distribution $\tilde{P}_n$ can be defined also for RV's $X_1, \cdots, X_n$ taking values in an arbitrary (rather than finite) set $\mathcal{X}$, cf. Section VII, (VII.21). Theorem III.3 and its extensions to arbitrary $\mathcal{X}$ are referred to as Sanov's theorem, the first general result being that of Sanov [72], cf. also Dembo and Zeitouni [39].

iii) The Corollary is an instance of "Gibbs' conditioning principle," cf. [39].

Proof: By (III.19) we have

$$\Pr\{\tilde{P}_n \in \Pi\} = \sum_{P \in \Pi \cap \mathcal{P}_n(\mathcal{X})} Q^n(\mathcal{T}_P^n). \quad \text{(III.24)}$$

By Lemma II.2 and (III.5), this gives

$$\begin{aligned} &|\mathcal{P}_n(\mathcal{X})|^{-1} \exp\{-nD(\Pi \cap \mathcal{P}_n(\mathcal{X})\|Q)\} \\ &\quad \le \Pr\{\tilde{P}_n \in \Pi\} \\ &\quad \le |\mathcal{P}_n(\mathcal{X})| \exp\{-nD(\Pi \cap \mathcal{P}_n(\mathcal{X})\|Q)\} \end{aligned} \quad \text{(III.25)}$$

whence Theorem III.3 follows.

To prove the Corollary, notice first that for any type $P \in \mathcal{P}_n(\mathcal{X})$ the conditional probability of $X_1 = a_1, \cdots, X_k = a_k$ on the condition $\tilde{P}_n = P$ is the same as the probability of the following: given an urn containing n balls marked with symbols $a \in \mathcal{X}$, where the number of balls marked by a equals $nP(a)$, if k balls are drawn without replacement their consecutive marks will be $a_1, \cdots, a_k$. Clearly, this probability approaches that for drawing with replacement, uniformly in the sense that the difference is less than any fixed $\varepsilon > 0$ for $n \geq n_0$ sufficiently large depending on k and ε only. Thus

$$\left| \Pr\{X_1 = a_1, \cdots, X_k = a_k | \tilde{P}_n = P\} - \prod_{i=1}^{k} P(a_i) \right| < \varepsilon$$

if

$$n \geq n_0(k, \varepsilon). \tag{III.26}$$

Let

$$U_\delta = \{P\colon |P(a) - P^*(a)| < \delta \quad \text{for each } a \in \mathcal{X}\} \tag{III.27}$$

be a small neighborhood of P^*. As $\overline{\Pi} \cap U_\delta^c$ is closed there exists $P^{**} \in \overline{\Pi} \cap U_\delta^c$ with $D(P^{**}||Q) = D(\overline{\Pi} \cap U_\delta^c||Q)$, and by the assumed uniqueness of P^* this implies

$$D(\overline{\Pi} \cap U_\delta^c||Q) = D(P^*||Q) + \eta, \qquad \text{for some } \eta > 0. \tag{III.28}$$

The Corollary follows since

$$\begin{aligned} &\left| \Pr\{X_1 = a_1, \cdots, X_k = a_k | \tilde{P}_n \in \Pi\} - \prod_{i=1}^{k} P^*(a_i) \right| \\ &\quad \leq \sum_{P \in \Pi \cap \mathcal{P}_n} |\Pr\{X_1 = a_1, \cdots, X_k = a_k | \tilde{P}_n = P\} \\ &\qquad - \prod_{i=1}^{k} P^*(a_i)| \Pr\{\tilde{P}_n = P | \tilde{P}_n \in \Pi\} \end{aligned}$$

where for sufficiently large n the absolute value term is arbitrarily small if $P \in \Pi \cap U_\delta$ with δ small, by (III.26) and (III.27), while the conditional probability factor is less than $\exp(-n\eta/2)$ if $P \in \Pi \cap U_\delta^c$, by (III.28), Lemma II.2, and (III.22).

IV. Error Exponents for DMC's

A first major success of the method of types was to gain better insight into the error exponent problem for DMC's. Theorem IV.1, below, due to Csiszár, Körner, and Marton [32], shows that the "random coding" error exponent for constant composition codes is attainable universally, i.e., by the same encoder and decoder, for all DMC's for which it is positive. An important feature of the proof is that, rather than bounding the expectation of the error probability for an ensemble of codes (and conclude that some code in the ensemble meets the obtained bound), the error probability is bounded for a given codeword set and a given decoder. The role of "random coding" reduces to show the existence of a "good" codeword set. We will also reproduce the simple "method of types" proof of the "sphere packing" bound for constant composition codes (Theorem IV.2, cf. Csiszár and Körner [30, p. 181]) establishing that the previous universally attainable error exponent is often (though not always) the best possible even for a known channel. This optimality holds among codes with a fixed codeword type, while the type yielding the best exponent depends on the actual channel.

Lemma IV.1: For any $R > 0$ and type $P \in \mathcal{P}_n(\mathcal{X})$, there exist $N \approx \exp(nR)$ sequences $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ in $\mathcal{T}_P^n$ such that for every joint type $P_{X\tilde{X}} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X})$ with $P_X = P_{\tilde{X}} = P$

$$|\{j\colon j \neq i, \quad (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}_{X\tilde{X}}^n\}| < \exp\{n(R - I(X \wedge \tilde{X}))\}, \quad i = 1, \cdots, N \tag{IV.1}$$

(cf. Section II for notation).

Remark: Equation (IV.1) implies that $I(\boldsymbol{x}_i \wedge \boldsymbol{x}_j) \leq R$ for each $i \neq j$. In particular, $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ are distinct sequences if $R \leq H(P)$.

Proof: By a simple random coding argument. For completeness, the proof will be given in the Appendix.

Theorem IV.1: For any DMC $\{W\colon \mathcal{X} \to \mathcal{Y}\}$, a code with codeword set $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ as in Lemma IV.1 and decoder $\varphi\colon \mathcal{Y}^n \to \{0, 1, \cdots, N\}$ defined by

$$\varphi(\boldsymbol{y}) = \begin{cases} i, & \text{if } I(\boldsymbol{x}_i \wedge \boldsymbol{y}) > I(\boldsymbol{x}_j \wedge \boldsymbol{y}) \text{ for each } j \neq i \\ 0, & \text{if no such } i \text{ exists} \end{cases} \tag{IV.2}$$

has maximum probability of error satisfying

$$\max_{1 \leq i \leq N} e_i \lesssim \exp\{-nE_r(R, P, W)\} \tag{IV.3}$$

where

$$e_i = W^n(\{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \neq i\}|\boldsymbol{x}_i) \tag{IV.4}$$

and

$$E_r(R, P, W) = \min_{\substack{P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \\ P_X = P}} [D(P_{XY}||W) + |I(X \wedge Y) - R|^+] \tag{IV.5}$$

is the "random coding" exponent for codeword type P.

Remarks:

i) Denote by $I(P, W)$ the mutual information $I(X \wedge Y)$ when $P_{XY} = P \times W$. Clearly, $E_r(R, P, W) > 0$ iff $R < I(P, W)$. Thus the channel-independent codes in Theorem IV.1 have exponentially decreasing error probability for every DMC with $I(P, W) > R$; it is well known that for channels with $I(P, W) < R$ no rate-R codes with codewords in $\mathcal{T}_P^n$ have small probability of error. The exponent $E_r(R, P, W)$ is best possible for many channels, namely, for those that satisfy $R_{cr}(P, W) \leq R < I(P, W)$, cf. Remark ii) to Theorem IV.2.

ii) It follows by a simple continuity argument that for any DMC

$$E_r(R, W) = \max_{P \in \mathcal{P}(\mathcal{X})} E_r(R, P, W)$$

is also an attainable error exponent with rate-R codes, though no longer channel-independent ones (as the maximizing P depends on W). This exponent is positive for every R less than the channel capacity $C(W) = \max_P I(P, W)$. It was first derived by Fano [44], and then in a simple way by Gallager [46], in algebraic forms different from that given above.

iii) The "empirical mutual information decoder" (IV.2) was first suggested by Goppa [48] as one not depending on the channel and still suitable to attain channel capacity. This decoder could be equivalently defined, perhaps more intuitively, by minimizing the "entropy distance" $H(\boldsymbol{x}_j|\boldsymbol{y})$ of the codewords from the received $\boldsymbol{y}$. Lemma IV.1 may be visualized as one asserting the existence of codes with good entropy distance distribution. As Blahut [16] observed, among sequences in a single type class the entropy distance satisfies the axioms of a metric, except that it may vanish for certain pairs of distinct sequences.

Proof of Theorem IV.1: As $W^n(\boldsymbol{y}|\boldsymbol{x}_i)$ is constant for $\boldsymbol{y} \in \mathcal{T}^n_{Y|X}(\boldsymbol{x}_i)$, we may write, using (II.7)

$$\begin{aligned} e_i &= \sum W^n(\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i) \cap \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \neq i\}|\boldsymbol{x}_i) \\ &\approx \sum \frac{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i) \cap \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \neq i\}|}{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i)|} \\ &\quad \cdot \exp\{-nD(P_{XY}||W)\} \end{aligned} \tag{IV.6}$$

where the summation is for all joint types $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ with $P_X = P$.

This and other sums below will be bounded in the $\lesssim$ sense by the largest term, without explicitly saying so.

To bound the cardinality ratios in (IV.6), notice that by the definition (IV.2) of φ, $\varphi(\boldsymbol{y}) \neq i$ iff there exists $j \neq i$ such that the joint type of $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{y})$ is represented by dummy RV's $X\tilde{X}Y$ with

$$I(X \wedge Y) \leq I(\tilde{X} \wedge Y). \tag{IV.7}$$

Thus, denoting by $\Pi_n(XY)$ the set of joint types

$$P_{X\tilde{X}Y} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X} \times \mathcal{Y})$$

with given P_{XY}, with $P_{\tilde{X}} = P$, and satisfying (IV.7),

$$\begin{aligned} &\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i) \cap \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \neq i\} \\ &\qquad = \bigcup_{P_{X\tilde{X}Y} \in \Pi_n(XY)} \bigcup_{j\colon j \neq i} \mathcal{T}^n_{Y|X\tilde{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j). \end{aligned} \tag{IV.8}$$

By Lemma II.3

$$\begin{aligned} \frac{|\mathcal{T}^n_{Y|X\tilde{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j)|}{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i)|} &\approx \frac{\exp\{nH(Y|X\tilde{X})\}}{\exp\{nH(Y|X)\}} \\ &= \exp\{-nI(\tilde{X} \wedge Y|X)\} \\ &= \exp\{-n[I(\tilde{X} \wedge XY) - I(X \wedge \tilde{X})]\} \\ &\leq \exp\{-n[I(\tilde{X} \wedge Y) - I(X \wedge \tilde{X})]\}. \end{aligned} \tag{IV.9}$$

Since $\mathcal{T}^n_{Y|X\tilde{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \neq \emptyset$ only when $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}^n_{X\tilde{X}}$, (IV.1) and (IV.9) imply that

$$\begin{aligned} &\frac{\left|\bigcup_{j\neq i} \mathcal{T}^n_{Y|X\tilde{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right|}{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i)|} \\ &\qquad \lesssim \sum_{\substack{P_{X\tilde{X}} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{X}) \\ P_X = P_{\tilde{X}} = P}} |\{j\colon j \neq i,\, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}^n_{X\tilde{X}}\}| \\ &\qquad\quad \cdot \exp\{-n[I(\tilde{X} \wedge Y) - I(X \wedge \tilde{X})]\} \\ &\qquad \lesssim \exp\{-n[I(\tilde{X} \wedge Y) - R]\}. \end{aligned} \tag{IV.10}$$

This bound remains valid if $I(\tilde{X} \wedge Y) - R$ is replaced by $|I(\tilde{X} \wedge Y) - R|^+$, since the left-hand side is always ≤ 1. Hence (IV.8) gives

$$\begin{aligned} &\frac{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i) \cap \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \neq i\}|}{|\mathcal{T}^n_{Y|X}(\boldsymbol{x}_i)|} \\ &\qquad \lesssim \exp\left\{-n \min_{P_{X\tilde{X}Y} \in \Pi_n(XY)} |I(\tilde{X} \wedge Y) - R|^+\right\} \\ &\qquad \leq \exp\{-n|I(X \wedge Y) - R|^+\} \end{aligned} \tag{IV.11}$$

where the last inequality holds by (IV.7). Substituting (IV.11) into (IV.6) gives (IV.3), completing the proof.

Theorem IV.2: Given arbitrary $R > 0$, $\delta > 0$, and DMC $\{W\colon \mathcal{X} \to \mathcal{Y}\}$, every code of sufficiently large blocklength n with $N \geq \exp\{n(R+\delta)\}$ codewords, each of the same type P, and with arbitrary decoder φ, has average probability of error

$$\frac{1}{N} \sum_{i=1}^{N} e_i \geq \exp\{-n[E_{sp}(R, P, W) + \delta]\} \tag{IV.12}$$

where

$$E_{sp}(R, P, W) = \min_{\substack{P_{XY} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \\ P_X = P,\, I(X\wedge Y) \leq R}} D(P_{XY}||W) \tag{IV.13}$$

is the "sphere packing" exponent for codeword type P.

Remarks:

i) It follows from Theorem IV.2 that even if the codewords are not required to be of the same type, the lower bound (IV.12) to the average probability of error always holds with $E_{sp}(R, P, W)$ replaced by

$$E_{sp}(R, W) = \max_P E_{sp}(R, P, W).$$

The latter is equal to the exponent in the "sphere packing bound" first proved by Shannon, Gallager, and Berlekamp [74]. The first simple proof of the sphere packing bound was given by Haroutunian [52]. The author is indebted to Haroutunian for the information that he had been aware also of the proof reproduced below but published only his other proof because it was not restricted to the discrete case.

ii) Both $E_{sp}(R, P, W)$ and $E_r(R, P, W)$ are convex functions of R, positive in the same interval $[0, I(P, W))$; they are equal if $R \geq R_{cr}(P, W)$ where

$R_{cr}(P, W)$ is the abscissa of the leftmost point where the graph of $E_{sp}(R, P, W)$ as a functions of R meets its supporting line of slope -1. The same holds for $E_{sp}(R, W)$ and $E_r(R, W)$ which are (positive and) equal in an interval $[R_{cr}(W), C(W))$. For R in this interval, their common value is the exact error exponent for rate-R codes. For smaller rates, the exact error exponent is still unknown.

iii) Dueck and Körner [42] showed that for codes with codeword set $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \subset \mathcal{T}_P^n$ and rate $(1/n)\log N > R + \delta$ with $R > I(P, W)$, the average probability of correct decoding goes to zero exponentially with exponent not smaller than the minimum of $D(P_{XY}\|W) + |R - I(X \wedge Y)|^+$ for $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ satisfying $P_X = P$. This result follows from (IV.16) by Lemma II.3. In [42] also its tightness was proved.

Proof of Theorem IV.2: Given a codeword set

$$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \subset \mathcal{T}_P^n$$

and arbitrary decoder φ, write

$$D_i = \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) = i\}, \qquad i = 1, \cdots, N. \tag{IV.14}$$

For every joint type $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ with $P_X = P$

$$\sum_{i=1}^{N} |\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i) \cap D_i| \le \left|\bigcup_{i=1}^{N} \mathcal{T}_{Y|X}^n(\boldsymbol{x}_i)\right| \le |\mathcal{T}_Y^n|. \tag{IV.15}$$

Hence, supposing $N \ge \exp\{n(R+\delta)\}$, it follows by (II.6) and (II.7) that

$$\frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i) \cap D_i|}{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i)|} \lesssim \frac{1}{N}\frac{\exp\{nH(Y)\}}{\exp\{nH(Y|X)\}} \le \exp\{n(I(X \wedge Y) - R - \delta)\}. \tag{IV.16}$$

In particular, if $I(X \wedge Y) \le R + \delta/2$ and $n \ge n(\delta)$ (sufficiently large), the left-hand side of (IV.16) is less than $1/2$, say, and hence

$$\frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i) \cap D_i^c|}{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i)|} \ge \frac{1}{2}, \quad \text{if } I(X \wedge Y) \le R + \frac{\delta}{2}, \ \ n \ge n_0(\delta). \tag{IV.17}$$

On account of (IV.4) and (IV.14), it follows from (IV.17) and Lemma II.3 that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} e_i &= \frac{1}{N}\sum_{i=1}^{N} W^n(D_i^c|\boldsymbol{x}_i) \\ &\ge \frac{1}{N}\sum_{i=1}^{N} W^n(\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i) \cap D_i^c|\boldsymbol{x}_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i) \cap D_i^c|}{|\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i)|} W^n(\mathcal{T}_{Y|X}^n(\boldsymbol{x}_i)|\boldsymbol{x}_i) \\ &\approx \exp\{-nD(P_{XY}\|W)\}\end{aligned} \tag{IV.18}$$

for joint types $P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ with $I(X \wedge Y) \le R + \delta/2$. A simple continuity argument shows that for sufficiently large n the minimum of $D(P_{XY}\|W)$ for these joint types is less than the minimum in (IV.13) plus δ, and this completes the proof.

Related Further Results

The proof technique of Theorem IV.1 has lead to various further results. Already in Csiszár, Körner, and Marton [32] a stronger result than Theorem IV.1 was proved, with an exponent better for "small rates" than $E_r(R, P, W)$. With the channel-dependent maximum-likehood decoder, a similar derivation yields an even better exponent for small rates that, when optimized with respect to the codeword type, gives the exponent of Gallager's [46] "expurgated bound" (cf. [30, pp. 185, 193]). In [32] the problem of separately bounding the erasure and undetected error probabilities was also addressed; a decoder $\varphi\colon \mathcal{Y}^n \to \{0, 1, \cdots, N\}$ yields an erasure if $\varphi(\boldsymbol{y}) = 0$, while an undetected error occurs if $\varphi(\boldsymbol{y})$ equals a message index $1 \le i \le N$ but not the correct one. Using a (still channel-independent) modification of the decoder (IV.2), jointly attainable exponents for erasure and undetected error probabilities were obtained (cf. [30, pp. 174–177]). Csiszár and Körner [28] derived exponential error bounds attainable with (a codeword set as in Lemma IV.1 and) decoders defined similarly to (IV.2) but with $\alpha(P_{\boldsymbol{xy}})$ instead of $I(\boldsymbol{x} \wedge \boldsymbol{y})$ for an arbitrary function α on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, possibly channel-dependent. Recently Telatar and Gallager [77] used channel-dependent decoders of a somewhat more general kind to derive jointly attainable exponents for erasure and undetected error probabilities improving upon those in [32].

A particular class of α-decoders received considerable attention recently. They are the d-decoders defined by minimizing a "distance"

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i), \qquad d(x, y) \text{ a given function on } \mathcal{X} \times \mathcal{Y} \tag{IV.19}$$

setting $\varphi(\boldsymbol{y}) = i$ if $d(\boldsymbol{x}_i, \boldsymbol{y}) < d(\boldsymbol{x}_j, \boldsymbol{y})$ for all $j \ne i$, and $\varphi(\boldsymbol{y}) = 0$ if no such i exists. Here the term "distance" is used in the widest sense, no restriction on d is implied. In this context, even the capacity problem is open, in general. The d-capacity of a DMC is the supremum of rates of codes with a given d-decoder that yield arbitrarily small probability of error. In the special case when $d(x, y) = 0$ or 1 according as $W(y|x) > 0$ or $= 0$, d-capacity provides the "zero undetected error" capacity or "erasures-only" capacity. Shannon's zero-error capacity can also be regarded as a special case of d-capacity, and so can the graph-theoretic concept of Sperner capacity, cf. [35].

A lower bound to d-capacity follows as a special case of a result in [28]; this bound was obtained also by Hui [55]. Balakirsky [12] proved by delicate "type" arguments the tightness of that bound for channels with binary input alphabet. Csiszár and Narayan [35] showed that the mentioned bound is not tight in general but its positivity is necessary for positive d-capacity. Lapidoth [59] showed that d-capacity can

equal the channel capacity $C(W)$ even if the above lower bound is strictly smaller. Other recent works addressing the problem of d-capacity or its special case of zero undetected error capacity include Merhav, Kaplan, Lapidoth, and Shamai [67], Ahlswede, Cai, and Zhang [9], as well as Telatar and Gallager [77].

V. Capacity of Arbitrarily Varying Channels

AVC's were introduced by Blackwell, Breiman, and Thomasian [15] to model communication situations where the channel statistics ("state") may vary in an unknown and arbitrary manner during the transmission of a codeword, perhaps caused by jamming. Formally, an AVC with (finite) input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and set of possible states $\mathcal{S}$ is defined by the probabilities $W(y|x, s)$ of receiving $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ is sent and $s \in \mathcal{S}$ is the state. The corresponding probabilities for n-length sequences are

$$W^n(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}) = \prod_{i=1}^{n} W(y_i|x_i, s_i). \tag{V.1}$$

The capacity problem for AVC's has many variants according to sender's and receivers's knowledge about the states, the state selector's knowledge about the codeword, degree of randomization in encoding and decoding, etc. Here we concentrate on the situation when no information is available to the sender and receiver about the states, nor to the state selector about the codeword sent, and only *deterministic codes* are permissible.

For a code with codeword set $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ and decoder $\varphi\colon \mathcal{Y}^n \to \{0, 1, \cdots, N\}$, the maximum and average probability of error are defined as

$$e = \max_{\boldsymbol{s}\in\mathcal{S}^n} \max_{1\le i\le N} e_i(\boldsymbol{s}) \qquad \overline{e} = \max_{\boldsymbol{s}\in\mathcal{S}^n} \frac{1}{N}\sum_{i=1}^{N} e_i(\boldsymbol{s}) \tag{V.2}$$

where

$$e_i(\boldsymbol{s}) = W^n(\{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \ne i\}|\boldsymbol{x}_i, \boldsymbol{s}). \tag{V.3}$$

The supremum of rates at which transmission with arbitrarily small maximum or average probability of error is possible, is called the m-capacity C_m or a-capacity C_a, respectively. Unlike for a DMC, $C_m < C_a$ is possible, and C_a may be zero when the "random coding capacity" C_r is positive. Here C_r is the supremum of rates for which ensembles of codes exists such that the expectation of $e_i(\boldsymbol{s})$ over the ensemble is arbitrarily small for $i = 1, \cdots, N$ and all $\boldsymbol{s} \in \mathcal{S}^n$. Already Blackwell, Breiman, and Thomasian [15] showed that

$$C_r = \max_{P_X\in\mathcal{P}(\mathcal{X})} \min_{P_S\in\mathcal{P}(\mathcal{S})} I(X \wedge Y) \quad \text{where } P_{XSY} = P_X \times P_S \times W \tag{V.4}$$

and gave an example where $C_r > C_a = 0$.

Below we review the presently available best result about m-capacity (Theorem V.1; Csiszár and Körner [29]), and the single-letter characterization of a-capacity (Theorem V.2; Csiszár and Narayan [33]). Both follow the pattern of Theorem IV.1: for a "good" codeword set and "good" decoder, the error probability is bounded via the method of types. A remarkable feature is that the very error-bounding process naturally suggests a good decoder.

Given an AVC defined by

$$\{W(y|x, s), x \in \mathcal{X}, s \in \mathcal{S}, y \in \mathcal{Y}\}$$

as above, for input symbols x and $\tilde{x}$ we write $x \sim \tilde{x}$ if there exists PD's Q and $\tilde{Q}$ on $\mathcal{S}$ such that

$$\sum_{s\in\mathcal{S}} W(y|x, s)Q(s) = \sum_{s\in\mathcal{S}} W(y|\tilde{x}, s)\tilde{Q}(s), \qquad \text{for all } y \in \mathcal{Y}. \tag{V.5}$$

The AVC is symmetrizable if there exists a channel $U\colon \mathcal{X} \to \mathcal{S}$ such that

$$\sum_{s\in\mathcal{S}} W(y|x, s)U(s|\tilde{x}) = \sum_{s\in\mathcal{S}} W(y|\tilde{x}, s)U(s|x) \quad \text{for all } x, \tilde{x} \in \mathcal{X},\ y \in \mathcal{Y}. \tag{V.6}$$

It has long been known that $C_m = 0$ iff $x \sim \tilde{x}$ for all $x, \tilde{x}$ in $\mathcal{X}$ [57], and that the right-hand side of (V.9) below is always an upper bound to C_m [10]. Ericson [43] showed that symmetrizability implies $C_a = 0$.

Theorem V.1: For $P \in \mathcal{P}(\mathcal{X})$ write

$$C(P) = \min_{P_{XS}\in\mathcal{P}(\mathcal{X}\times\mathcal{S}),\, P_X = P} I(X \wedge Y) \quad \text{where } P_{XSY} = P_{XS} \times W \tag{V.7}$$

and

$$D(P) = \min I(X \wedge \tilde{X}) \quad \text{subject to } P_X = P_{\tilde{X}} = P, \Pr\{X \sim \tilde{X}\} = 1. \tag{V.8}$$

Then $\min[C(P), D(P)]$ is an achievable rate for the maximum probability of error criterion, for each $P \in \mathcal{P}(\mathcal{X})$. Hence

$$C_m = \max_{P\in\mathcal{P}(\mathcal{X})} C(P) \tag{V.9}$$

if the maximum is attained for some P with $C(P) \le D(P)$.

Theorem V.2: For a nonsymmetrizable AVC, $C_a = C_r$.

Remarks: The first strong attack at a-capacity was that of Dobrushin and Stambler [40]. They were first to use large deviations arguments (including Chernoff bounding for dependent RV's) as well as "method of types" calculations to show that for randomly selected codes, the probability that $N^{-1}\sum_i e_i(\boldsymbol{s})$ is not small for a fixed $\boldsymbol{s} \in \mathcal{S}^n$ goes to zero doubly exponentially, implying the same also for $\overline{e}$. Unfortunately, as the method of types was not yet developed at that time, much effort had to be spent on technicalities. This diverted the authors' attention from what later turned out to be a key issue, viz., the choice of a good decoder, causing the results to fall short of a complete solution of the a-capacity problem. Not much later, Ahlswede [2] found a shortcut to that problem, proving by a clever trick that $C_a = C_r$ whenever $C_a > 0$; however, a single-letter necessary and sufficient condition for $C_a > 0$ remained elusive. Remarkably, the sufficiency of nonsymmetrizability for $C_a > 0$ does not seem

easier to prove than its sufficiency for $C_a = C_r$. The strongest result about m-capacity preceding [29] was that of Ahlswede [4] who proved (V.9) for AVC's such that $x \sim \tilde{x}$ never holds when $x \neq \tilde{x}$. He used large deviations arguments similar to those of Dobrushin and Stambler [40], and was first to use a sophisticated decoder (but not the one below). A full solution of the m-capacity problem appears a long way ahead; one of its special cases is Shannon's celebrated zero-error capacity problem.

Proof of Theorems V.1 and V.2: First one needs an analog of Lemma IV.1 about the existence of "good" codeword sets. This is the following

Lemma V.1: For any $\varepsilon > 0$, $R > \varepsilon$, and sufficiently large n, for each type $P \in \mathcal{P}_n(\mathcal{X})$ with $H(P) \geq R$ there exist $N \approx \exp(nR)$ sequences $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ in $\mathcal{T}_P^n$ such that for all joint types $P_{X\tilde{X}S} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X} \times \mathcal{Y})$ and all $\boldsymbol{s} \in \mathcal{S}^n$

$$|\{j\colon (\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s}) \in \mathcal{T}^n_{X\tilde{X}S}\}| \leq \exp\{n[|R - I(XS \wedge \tilde{X})|^+ + \varepsilon]\}, \quad i = 1, \cdots, N \tag{V.10}$$

$$\frac{1}{N}|\{i\colon I(\boldsymbol{x}_i \wedge \boldsymbol{s}) > \varepsilon\}| \leq \exp(-n\varepsilon/2) \tag{V.11}$$

$$\frac{1}{N}|\{i\colon (\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s}) \in \mathcal{T}^n_{X\tilde{X}S} \text{ for some } j \neq i\}| \leq \exp(-n\varepsilon/2), \quad \text{if } I(X \wedge \tilde{X}S) > |R - I(\tilde{X} \wedge S)|^+ + \varepsilon \tag{V.12}$$

$$I(\boldsymbol{x}_i \wedge \boldsymbol{x}_j) \leq R, \qquad \text{if } i \neq j. \tag{V.13}$$

This lemma can be proved by random selection, although more refined arguments are needed to get (V.10)–(V.12) than those in the proof of Lemma IV.1. One has to show that only with probability going to zero faster than exponentially will the randomly selected codewords violate (V.10)–(V.12), for any fixed $\boldsymbol{s} \in \mathcal{S}^n$, and i in (V.10). This can be done by Chernoff bounding. One difficulty is that in case of (V.12) dependent RV's have to be dealt with; this can be overcome using an idea of Dobrushin and Stambler [40]. For details, cf. [29] and [33].

We will use a codeword set as in Lemma V.1, with a decoder φ whose exact form will be suggested by the very error-bounding process.

Denote by $\Pi^m(\eta)$ and $\Pi^a(\eta)$ the family of those joint distributions $P_{XSY} \in \mathcal{P}(\mathcal{X} \times \mathcal{S} \times \mathcal{Y})$ that satisfy

$$D(P_{XSY}||P_{XS} \times W) < \eta, \qquad P_X = P \tag{V.14}$$

respectively,

$$D(P_{XSY}||P_X \times P_S \times W) < \eta, \qquad P_X = P. \tag{V.15}$$

Notice that $\Pi^a(\eta) \subset \Pi^m(\eta)$ since

$$D(P_{XSY}||P_X \times P_S \times W) = D(P_{XSY}||P_{XS} \times W) + I(X \wedge S). \tag{V.16}$$

A codeword $\boldsymbol{x}_i$ and a received sequence $\boldsymbol{y}$ may be considered "jointly typical" if there exists $\boldsymbol{s} \in \mathcal{S}^n$ such that the joint type of $(\boldsymbol{x}_i, \boldsymbol{s}, \boldsymbol{y})$ belongs to $\Pi^m(\eta)$ or $\Pi^a(\eta)$. The contribution to the maximum or average probability of error of output sequences $\boldsymbol{y}$ not jointly typical in this sense with the codeword sent is negligible. Indeed, we have by Lemmas II.1 and II.3, writing $\mathcal{P}_n$ as a shorthand for $\mathcal{P}_n(\mathcal{X} \times \mathcal{S} \times \mathcal{Y})$

$$W^n\left(\bigcup_{P_{XSY} \in \mathcal{P}_n \setminus \Pi^m(\eta)} \mathcal{T}^n_{Y|XS}(\boldsymbol{x}_i, \boldsymbol{s})|\boldsymbol{x}_i, \boldsymbol{s}\right) \lesssim \exp(-n\eta) \tag{V.17}$$

and—using also (V.16) and (V.11)—

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} W^n\left(\bigcup_{P_{XSY} \in \mathcal{P}_n \setminus \Pi^a(\eta)} \mathcal{T}^n_{Y|XS}(\boldsymbol{x}_i, \boldsymbol{s})|\boldsymbol{x}_i, \boldsymbol{s}\right) &\leq \frac{1}{N}|\{i\colon I(\boldsymbol{x}_i \wedge \boldsymbol{s}) > \varepsilon\}| + \frac{1}{N}\sum_{i\colon I(\boldsymbol{x}_i \wedge \boldsymbol{s}) \leq \varepsilon} \\ &\quad \cdot W^n\left(\bigcup_{P_{XSY} \in \mathcal{P}_n \setminus \Pi^m(\eta - \varepsilon)} \mathcal{T}_{Y|XS}(\boldsymbol{x}_i, \boldsymbol{s})|\boldsymbol{x}_i, \boldsymbol{s}\right) \\ &\lesssim \exp(-n\varepsilon/2) + \exp(-n(\eta - \varepsilon)).\end{aligned} \tag{V.18}$$

Motivated by (V.17) and (V.18) we will consider

$$L(\boldsymbol{y}) = \{i\colon P_{\boldsymbol{x}_i \boldsymbol{s} \boldsymbol{y}} \in \Pi \text{ for some } \boldsymbol{s} \in \mathcal{S}^n\} \tag{V.19}$$

as the list of candidates for the decoder output $\varphi(\boldsymbol{y})$, where Π denotes $\Pi^m(\eta)$ or $\Pi^a(\eta)$ according as the maximum or average probability of error criterion is being considered.

Dobrushin and Stambler [40] used, effectively, a decoder whose output was i if i was the only candidate in the above sense (with $\Pi = \Pi^a(\eta)$), while otherwise an error was declared. This "joint typicality decoder" has been shown suitable to achieve the a-capacity C_a for some but not all AVC's. To obtain a more powerful decoder, when the list (V.19) contains several messages one may reject some of them by a suitable rule. If only one $i \in L(\boldsymbol{y})$ remains unrejected, that will be the decoder output. We will consider rejection rules corresponding to sets $\Psi \subset \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y})$ of joint distributions $P_{X\tilde{X}SY}$ as follows: a candidate $i \in L(\boldsymbol{y})$ is rejected if for every $\boldsymbol{s} \in \mathcal{S}^n$ with $P_{\boldsymbol{x}_i \boldsymbol{s} \boldsymbol{y}} \in \Pi$ there exists $j \neq i$ in $L(\boldsymbol{y})$ such that the joint type of $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s}, \boldsymbol{y})$ belongs to Ψ. To reflect that $P_{\boldsymbol{x}_i \boldsymbol{s} \boldsymbol{y}} \in \Pi$ and that $P_{\boldsymbol{x}_j \tilde{\boldsymbol{s}} \boldsymbol{y}} \in \Pi$ for some $\tilde{\boldsymbol{s}} \in \mathcal{S}^n$ (as $j \in L(\boldsymbol{y})$), we assume that Ψ consists of such joint distributions $P_{X\tilde{X}SY}$ whose marginals P_{XSY} and $P_{\tilde{X}Y}$ satisfy

$$\begin{aligned}&P_{XSY} \in \Pi \\ &P_{\tilde{X}Y} \text{ is the marginal of some } P_{\tilde{X}\tilde{S}Y} \in \Pi.\end{aligned} \tag{V.20}$$

A set Ψ of joint distributions with the properties (V.20) will be called permissible if for each $\boldsymbol{y} \in \mathcal{Y}^n$, the rejection rule corresponding to Ψ leaves at most one $i \in L(\boldsymbol{y})$ unrejected. Then $\varphi(\boldsymbol{y})$ is set equal to the unrejected $i \in L(\boldsymbol{y})$, at 0 if no such i exists. For such a decoder φ, preliminarily with an unspecified $\Psi \subset \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y})$, the maximum or average probability of error can be bounded via standard "method of types" technique. The result will suggest a natural choice of Ψ that makes the bound exponentially small under the hypotheses of Theorems V.1 and V.2. The calculation is technical but instructive; it will be given in the Appendix.

Further Developments

By the above approach, Csiszár, and Narayan [33] also determined the a-capacity $C_a(\Lambda)$ for AVC's with state constraint Λ. The latter means that only those $\boldsymbol{s} \in \mathcal{S}^n$ are permissible state sequences that satisfy $n^{-1}\sum_i \ell(s_i) \le \Lambda$ for a given cost function ℓ on $\mathcal{S}$. For this case, Ahlswede's [2] trick does not work and, indeed, $0 < C_a(\Lambda) < C_r(\Lambda)$ may happen. Specializing the result to the (noiseless) binary adder AVC ($\mathcal{X} = \mathcal{S} = \mathcal{Y} = \{0, 1\}, W(y|x, s) = 1$ if $y = x + s \bmod 2$) with $\ell(s) = s$, the intuitive but nontrivial result was obtained that for $\Lambda < 1/2$, the a-capacity $C_a(\Lambda)$ equals the capacity of the binary symmetric DMC with crossover probability Λ. Notice that the corresponding m-capacity is the maximum rate of codes admitting correction of each pattern of no more than Λn bit errors, whose determination is a central open problem of coding theory. The role of the decoding rule was further studied, and a-capacity for some simple cases was explicitly calculated in Csiszár and Narayan [34].

While symmetrizable AVC's have $C_a = 0$, if the decoder output is not required to be a single message index i but a list of k candidates $i_1, \cdots, i_k$, the resulting "list code capacity" may be nonzero already for $k = 2$. Pinsker conjectured in 1990 that the "list-size-k" a-capacity is always equal to C_r if k is sufficiently large. Contributions to this problem, establishing Pinsker's conjecture, include Ahlswede and Cai [7], Blinovsky, Narayan, and Pinsker [17], and Hughes [54]. The last paper follows and extends the approach in this section. For AVC's with $C_r > 0$, a number M called the symmetrizability is determined, and the list-size-k a-capacity is shown to be 0 for $k \le M$ and equal to C_r for $k > M$. A partial analog of this result for list-size-k m-capacity is that the limit of the latter as $k \to \infty$ is always given by (V.9), cf. Ahlswede [6].

The approach in this section was extended to multiple-access AVC's by Gubner [50], although the full analog of Theorem V.2 was only conjectured by him. This conjecture was recently proved by Ahlswede and Cai [8]. Previously, Jahn [56] determined the a-capacity region of a multiple-access AVC under the condition that it had nonvoid interior, and showed that then the a-capacity region coincided with the random coding capacity region.

VI. Other Typical Applications

A. Rate-Distortion Theory

The usefulness for rate-distortion theory of partitioning n-length sequences into type classes was first demonstrated by Berger [13]. He established the following lemma, called the type covering lemma in [30].

Lemma VI.1: Given finite sets $\mathcal{X}, \mathcal{Y}$ and a nonnegative function $d(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, for $P \in \mathcal{P}_n(\mathcal{X})$ and $D \ge 0$ let $N_n(P, D)$ denote the minimum number of "d-balls of radius D"

$$B(\boldsymbol{y}, D) = \{\boldsymbol{x}\colon d(\boldsymbol{x}, \boldsymbol{y}) \le D\} \tag{VI.1}$$

needed to cover the type class $\mathcal{T}_P^n \subset \mathcal{X}^n$, where $d(\boldsymbol{x}, \boldsymbol{y})$ is defined by (IV.19). Then

$$\max_{P \in \mathcal{P}_n(\mathcal{X})} \left| \frac{1}{n} \log N_n(P, D) - R(P, D) \right| \to 0, \qquad \text{as } n \to \infty \tag{VI.2}$$

where

$$R(P, D) = \min I(X \wedge Y) \quad \text{subject to } P_X = P \quad Ed(X, Y) \le D \tag{VI.3}$$

is the "rate-distortion function."

This lemma is seen today as a consequence of a simple general result about coverings known as the Johnson–Stein–Lovász theorem, cf. Cohen *et al.* [21, p. 322]. The latter is useful in information theory also in other contexts, cf. Ahlswede [3].

An immediate consequence of Lemma VI.1 is that the minimum (asymptotic) rate of source block codes admitting the reproduction of each $\boldsymbol{x} \in \mathcal{X}^n$ by some $\boldsymbol{y} \in \mathcal{Y}^n$ with distortion $d(\boldsymbol{x}, \boldsymbol{y}) \le D$, is equal to $\max_{P \in \mathcal{P}(\mathcal{X})} R(P, D)$. Berger [13] also used this lemma to derive the rate-distortion theorem for arbitrarily varying sources.

As another application of Lemma VI.1, Marton [64] determined the error exponent for the compression of memoryless sources with a fidelity criterion. In fact, her error exponent is attainable universally (with codes depending on the "distortion measure" d but not on the source distribution Q or the distortion threshold D). Thus the following extension of Theorem III.1 holds, cf. [30, p. 156]: Given

$$0 < R < \max_{P \in \mathcal{P}(\mathcal{X})} R(P, D)$$

there exist codeword sets $C_n \subset \mathcal{Y}^n$ such that

$$\frac{1}{n} \log |C_n| \to R \tag{VI.4}$$

and for every PD $Q \in \mathcal{P}(\mathcal{X})$ and every $D \ge 0$

$$\frac{1}{n} \log Q^n \left(\left(\bigcup_{\boldsymbol{y} \in C_n} B(\boldsymbol{y}, D) \right)^c \right) \to -F(R, Q, D) \tag{VI.5}$$

where

$$F(R, Q, D) = \inf_{P\colon R(P, D) > R} D(P\|Q). \tag{VI.6}$$

Moreover, for any sequence of sets $C_n \subset \mathcal{Y}^n$ satisfying (VI.4) the liminf of the left-hand side of (VI.5) is $\ge -F(R, Q, D)$. Remarkably, the exponent (VI.6) is not necessarily a continuous function of R, cf. Ahlswede [5], although as Marton [64] showed, it is continuous when $d(\boldsymbol{x}, \boldsymbol{y})$ is the normalized Hamming distance (cf. also [30, p. 158]).

Recently, several papers have been devoted to the redundancy problem in rate-distortion theory, such as Yu and Speed [82], Linder, Lugosi, and Zeger [61], Merhav [65], Zhang, Yang, and Wei [83]. One version of the problem concerns the "rate redundancy" of D-semifaithful codes. A D-semifaithful code is defined by a mapping $f\colon \mathcal{X}^n \to \mathcal{Y}^n$ such that $d(\boldsymbol{x}, f(\boldsymbol{x})) \le D$ for all $\boldsymbol{x} \in \mathcal{X}^n$, together with an assignment to each $\boldsymbol{y}$ in the range of f of a binary codeword of length

$\ell(\boldsymbol{y})$, subject to prefix condition. Yu and Speed [82] showed that for a memoryless source with generic distribution Q there exist D-semifaithful codes whose rate redundancy

$$\frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} Q^n(\boldsymbol{x})\ell(f(\boldsymbol{x})) - R(Q, D) \tag{VI.7}$$

is less than a constant times $\log n/n$, moreover, such codes may be given universally (not depending on Q). They also conjectured that the rate redundancy (VI.7) can never be less than a constant times $\log n/n$, under a technical condition that excludes the case $D = 0$. Zhang, Yang, and Wei [83] proved that conjecture, and also determined the exact asymptotics of the "distortion redundancy" of the best rate-R block codes, again under some technical condition. This result of [83] says that for a memoryless source with generic distribution Q, the minimum of

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} Q^n(\boldsymbol{x})d(\boldsymbol{x}, f(\boldsymbol{x})) - D(Q, R) \tag{VI.8}$$

for mappings $f\colon \mathcal{X}^n \to C_n \subset \mathcal{Y}^n$ with $|\mathcal{C}_n| \le \exp(nR)$ is asymptotically equal to constant times $\log n/n$, with an explicitly given constant. Here $D(Q, R)$ is the inverse function of $R(Q, D)$ defined by (VI.3), with Q fixed. Both papers [82] and [83] heavily rely on the method of types. The latter represents one of the very few cases where the delicate calculations require more exact bounds on the cardinality and probability of a type class than the crude ones in Lemma II.2.

B. Source-Channel Error Exponent

When a memoryless source with alphabet $\mathcal{S}$ and generic distribution Q is transmitted over a DMC $\{W\colon \mathcal{X} \to \mathcal{Y}\}$ using a source-channel block code with encoder $f\colon \mathcal{S}^n \to \mathcal{X}^n$ and decoder $\varphi\colon \mathcal{Y}^n \to \mathcal{S}^n$, the probability of error is

$$\sum_{\boldsymbol{s}\in\mathcal{S}^n} Q^n(\boldsymbol{s})W^n(\{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) \ne \boldsymbol{s}\}|f(\boldsymbol{s})). \tag{VI.9}$$

Using techniques as in the proof of Theorem IV.1, Csiszár [22] showed that by suitable source-channel codes of block-length $n \to \infty$, not depending on Q, the error probability (VI.9) can be made exponentially small whenever $H(Q) < C(W)$, with exponent $\min_R[e_Q(R) + E_r(R, W)]$ (cf. (III.3) and the remarks to Theorems IV.1 and IV.2 for notation). This exponent is best possible if the minimum is attained for some $R \ge R_{cr}(W)$. For further results in this direction, including source-channel transmission with a distortion threshold, cf. Csiszár [24].

C. Multiterminal Source Coding

Historically, the first multiuser problem studied via the method of types was that of the error exponent for the Slepian–Wolf [76] problem, i.e., separate coding of (memoryless) correlated sources. Given a source pair with generic variables (X, Y), the error probability of an n-length block code with separate encoders f, g and common decoder φ is

$$\Pr\{\varphi(f(X^n), g(Y^n)) \ne (X^n, Y^n)\} \tag{VI.10}$$

where (X^n, Y^n) represents n independent repetitions of (X, Y). Csiszár, Körner, and Marton proved in 1977 (published in [27], cf. also [30, pp. 264–266]) that for suitable codes as above, with encoders that map $\mathcal{X}^n$ and $\mathcal{Y}^n$ into codeword sets of sizes

$$\|f\| \approx \exp(nR_1) \qquad \|g\| \approx \exp(nR_2), \tag{VI.11}$$

the error probability (VI.10) goes to zero exponentially as $n \to \infty$, with exponent

$$\begin{aligned} E_1(R_1, R_2, X, Y) = \min_{P_{\tilde{X}\tilde{Y}}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} [&D(P_{\tilde{X}\tilde{Y}}\|P_{XY}) \\ &+ |\min(R_1 - H(\tilde{X}|\tilde{Y}), R_2 - H(\tilde{Y}|\tilde{X}), \\ &\quad R_1 + R_2 - H(\tilde{X}\tilde{Y}))|^+] \end{aligned} \tag{VI.12}$$

whenever (R_1, R_2) is in the interior of the achievable rate region [76]

$$\begin{aligned} \mathcal{R}(X, Y) = \{(R_1, R_2)\colon &R_1 \ge H(X|Y), R_2 \ge H(Y|X), \\ &R_1 + R_2 \ge H(X, Y)\}. \end{aligned} \tag{VI.13}$$

This assertion can be proved letting φ be the "minimum entropy decoder" that outputs for any pair of codewords that pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ whose nonprobabilistic entropy $H(\boldsymbol{x}, \boldsymbol{y})$ is minimum among those having the given codewords (ties may be broken arbitrarily). Using this φ, the incorrectly decoded pairs $(\boldsymbol{x}, \boldsymbol{y})$ belong to one of the following three sets:

$$\begin{aligned} D_1 = \{&(\boldsymbol{x}, \boldsymbol{y})\colon H(\boldsymbol{x}', \boldsymbol{y}) \le H(\boldsymbol{x}, \boldsymbol{y}) \\ &\text{for some } \boldsymbol{x}' \ne \boldsymbol{x} \text{ with } f(\boldsymbol{x}') = f(\boldsymbol{x})\} \\ D_2 = \{&(\boldsymbol{x}, \boldsymbol{y})\colon H(\boldsymbol{x}, \boldsymbol{y}') \le H(\boldsymbol{x}, \boldsymbol{y}) \\ &\text{for some } \boldsymbol{y}' \ne \boldsymbol{y} \text{ with } g(\boldsymbol{y}') = g(\boldsymbol{y})\} \\ D_3 = \{&(\boldsymbol{x}, \boldsymbol{y})\colon H(\boldsymbol{x}', \boldsymbol{y}') \le H(\boldsymbol{x}, \boldsymbol{y}) \\ &\text{for some } \boldsymbol{x}' \ne \boldsymbol{x} \text{ and } \boldsymbol{y}' \ne \boldsymbol{y} \\ &\text{with } f(\boldsymbol{x}') = f(\boldsymbol{x}),\ g(\boldsymbol{y}') = g(\boldsymbol{y})\}. \end{aligned}$$

It can be seen by random selection that there exist f and g satisfying (VI.11) such that for each joint type $P_{\tilde{X}\tilde{Y}} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$

$$\begin{aligned} \frac{|D_1 \cap \mathcal{T}^n_{\tilde{X}\tilde{Y}}|}{|\mathcal{T}^n_{\tilde{X}\tilde{Y}}|} &\lesssim \exp\{-n|R_1 - H(\tilde{X}|\tilde{Y})|^+\} \\ \frac{|D_2 \cap \mathcal{T}^n_{\tilde{X}\tilde{Y}}|}{|\mathcal{T}^n_{\tilde{X}\tilde{Y}}|} &\lesssim \exp\{-n|R_2 - H(\tilde{Y}|\tilde{X})|^+\} \\ \frac{|D_3 \cap \mathcal{T}^n_{\tilde{X}\tilde{Y}}|}{|\mathcal{T}^n_{\tilde{X}\tilde{Y}}|} &\lesssim \exp\{-n|R_1 + R_2 - H(\tilde{X}\tilde{Y})|^+\}. \end{aligned}$$

Hence the assertion follows by Lemmas II.1 and II.2.

The error exponent (VI.12) for the Slepian–Wolf problem is attainable universally, i.e., with codes not depending on the distribution of (X, Y). This result is a counterpart of Theorem IV.1 for DMC's. The counterpart of Theorem IV.2 was also established by Csiszár, Körner, and Marton, *loc cit:* For no source pair can the error probability of codes satisfying (VI.11) decrease faster than with exponent

$$E_2(R_1, R_2, X, Y) = \min_{P_{\tilde{X}\tilde{Y}}\colon (R_1,R_2)\notin\mathcal{R}(\tilde{X},\tilde{Y})} D(P_{\tilde{X}\tilde{Y}}\|P_{XY}). \tag{VI.14}$$

The functions in (VI.12) and (VI.14) are equal if (R_1, R_2) is close to the boundary of $\mathcal{R}(X, Y)$. For such rate pairs, their common value gives the exact error exponent.

Remark: For the intrinsic relationship of source and channel problems, cf., e.g., Csiszár and Körner [30, Secs. III.1 and III.2], Csiszár and Körner [28], and Ahlswede [3, Part II].

The above results have been extended in various ways. Extensions to more than two correlated sources are straightforward, cf. [27], or [30, pp. 267, 268]. Csiszár and Körner [28] showed that good encoders can be obtained, instead of random selection, also by a graph-theoretic approach. Another contribution of [28], in retrospect the more important one, was to apply the method of types to study the performance of various decoders, and to improve the exponent (VI.12) for "large" rates. Csiszár [23] showed that the exponent (VI.12) is (universally) attainable also with linear codes, i.e., constraining f and g be linear maps (to this, $\mathcal{X}$ and $\mathcal{Y}$ have to be fields, but that can always be assumed, extending the alphabets by dummy symbols of zero probability if necessary). Also in [23], linear codes were shown to give better than the previously known best exponent for certain rate pairs. More recently, Oohama and Han [70] obtained another improvement for certain rate pairs, and Oohama [69] determined the exact exponent for a modified version of the problem. That modification admits partial cooperation of the encoders, which, however, does not affect the achievable rate region (VI.13) nor the upper bound (VI.14) to achievable error exponents. On the other hand, the modification makes the exponent in (VI.14) achievable for all rate pairs (R_1, R_2) in the interior of $\mathcal{R}(X, Y)$, even universally.

D. Multiterminal Channels

The first application of the method of types to a multiterminal channel coding problem was the paper of Körner and Sgarro [58]. Using the same idea as in Theorem IV.1, they derived an error exponent for the asymmetric broadcast channel, cf. [30, p. 359] for the definition of this channel.

Here let us concentrate on the multiple-access channel (MAC). A MAC with input alphabets $\mathcal{X}$, $\mathcal{Y}$, and output alphabet $\mathcal{Z}$ is formally a DMC $\{W\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}\}$, with the understanding that there are two (noncommunicating) senders, one selecting the $\mathcal{X}$-component the other the $\mathcal{Y}$-component of the input. Thus codes with two codewords sets $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\} \subset \mathcal{X}^n$ and $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N\} \subset \mathcal{Y}^n$ are considered, the decoder φ assigns a pair of message indices i, j to each $\boldsymbol{z} \in \mathcal{Z}^n$, and the average probability of error is

$$\overline{e} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} W^n(\{\boldsymbol{z}\colon \varphi(\boldsymbol{z}) \neq (i, j)\} | \boldsymbol{x}_i, \boldsymbol{y}_j). \quad \text{(VI.15)}$$

The capacity region, i.e., the closure $\mathcal{C}$ of the set of those rate pairs (R_1, R_2) to which codes with $M \approx \exp nR_1$, $N \approx \exp nR_2$, and $\overline{e} \to 0$ exist, is characterized as follows: $(R_1, R_2) \in \mathcal{C}$ iff there exist RV's U, X, Y, Z, with U taking values in an auxiliary set $\mathcal{U}$ of size $|\mathcal{U}| = 2$, whose joint distribution is of form

$$P_{UXYZ}(u, x, y, z) = P_U(u)P_{X|U}(x|u)P_{Y|U}(y|u)W(z|x, y) \quad \text{(VI.16)}$$

and such that

$$\begin{aligned} R_1 &\leq I(X \wedge Z|YU) \\ R_2 &\leq I(Y \wedge Z|XU) \\ R_1 + R_2 &\leq I(XY \wedge Z|U). \end{aligned} \quad \text{(VI.17)}$$

The capacity region $\mathcal{C}$ was first determined (in a different algebraic form) by Ahlswede [1] and Liao [60]. The maximum probability of error criterion may give a smaller region, cf. Dueck [41]; a single-letter characterization of the latter is not known.

For (R_1, R_2) in the interior of $\mathcal{C}$, $\overline{e}$ can be made exponentially small; Gallager [47] gave an attainable exponent everywhere positive in the interior of $\mathcal{C}$. Pokorny and Wallmeier [71] were first to apply the method of types to this problem. They showed the existence of (universal) codes with codewords $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ whose joint types with a fixed $\boldsymbol{u} \in \mathcal{U}^n$ are arbitrarily given $P_{UX} \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X})$ and $P_{UY} \in \mathcal{P}_n(\mathcal{U} \times \mathcal{Y})$ such that the average probability of error is bounded above exponentially, with exponent depending on P_{UX}, P_{UY}, and W; that exponent is positive for each W with the property that for P_{UXYZ} determined by (VI.16) with the given P_{UX} and P_{UY}, (VI.17) is satisfied with strict inequalities. Pokorny and Wallmeier [71] used the proof technique of Theorem IV.1 with a decoder maximizing $I(\boldsymbol{x}_i\boldsymbol{y}_j \wedge \boldsymbol{z}|\boldsymbol{u})$. Recently, Liu and Hughes [62] improved upon the exponent of [71], using a similar technique but with decoder minimizing $H(\boldsymbol{x}_i\boldsymbol{y}_j|\boldsymbol{z}\boldsymbol{u})$. The "maximum mutual information" and "minimum conditional entropy" decoding rules are equivalent for DMC's with codewords of the same type but not in the MAC context; by the result of [62], "minimum conditional entropy" appears the better one.

VII. Extensions

While the type concept is originally tailored to memoryless models, it has extensions suitable for more complex models, as well. So far, such extensions proved useful mainly in the context of source coding and hypothesis testing.

Abstractly, given any family of source models with alphabet $\mathcal{X}$, a partition of $\mathcal{X}^n$ into sets $A_1, \cdots, A_{N_n}$ can be regarded as a partition into "type classes" if sequences in the same A_i are equiprobable under each model in the family. Of course, a subexponential growth rate of N_n is desirable. This general concept can be applied, e.g., to variable-length universal source coding: assign to each $\boldsymbol{x} \in A_i$ a binary codeword of length $\ell(\boldsymbol{x}) = \lceil \log_2 N_n \rceil + \lceil \log_2 |A_i| \rceil$, the first $\lceil \log_2 N_n \rceil$ bits specifying the class index i, the last $\lceil \log_2 |A_i| \rceil$ bits identifying $\boldsymbol{x}$ within A_i. Clearly, $\ell(\boldsymbol{x})$ will exceed the "ideal codelength" $-\log_2 P(\boldsymbol{x})$ by less that $\lceil \log_2 N_n \rceil + 1$, for each source model in the family.

As an example, consider the model family $\mathcal{R}$ of renewal processes, i.e., binary sources such that the lengths of 0 runs preceeding the 1's are i.i.d. RV's. Define the renewal type of a sequence $\boldsymbol{x} \in \{0, 1\}^n$ as $(k_0, k_1, \cdots)$ where k_j denotes

the number of 1's in $\boldsymbol{x}$ which are preceeded by exactly j consecutive 0's. Sequences $\boldsymbol{x} \in \{0, 1\}^n$ of the same renewal type are equiprobable under each model in $\mathcal{R}$, and renewal types can be used for the model family $\mathcal{R}$ much in the same way as standard types for memoryless sources. Csiszár and Shields [37] showed that there are $N_n = \exp(O(\sqrt{n}))$ renewal type classes, which implies the possibility of universal coding with redundancy $O(\sqrt{n})$ for the family of renewal processes. It was also shown in [37] that the redundancy bound $O(\sqrt{n})$ is best possible for this family, as opposed to finitely parametrizable model families for which the best attainable redundancy is typically $O(\log n)$.

Below we briefly discuss some more direct extensions of the standard type concept.

A. Second- and Higher Order Types

The type concept appropriate for Markov chains is "second-order type," defined for a sequence $\boldsymbol{x} = x_1 \cdots x_n \in \mathcal{X}^n$ as the PD $P_{\boldsymbol{x}}^{(2)} \in \mathcal{P}_{n-1}(\mathcal{X}^2)$ with

$$P_{\boldsymbol{x}}^{(2)}(a, b) = \frac{1}{n-1} |\{i\colon x_i = a, x_{i+1} = b\}|. \tag{VII.1}$$

In other words, $P_{\boldsymbol{x}}^{(2)}$ is the joint type of $\boldsymbol{x}' = x_1 \cdots x_{n-1}$ and $\boldsymbol{x}'' = x_2 \cdots x_n$. Denote by $\mathcal{P}_n^{(2)}(\mathcal{X}, a)$ the set of all possible second-order types of sequences $\boldsymbol{x} \in \mathcal{X}^n$ with $x_1 = a$, and for dummy RV's X, Y representing such a second-order type (i.e., $P_{XY} \in \mathcal{P}_n^{(2)}(\mathcal{X}, a)$) let $\mathcal{T}_{XY,a}^{n,2}$ denote the type class $\{\boldsymbol{x}\colon \boldsymbol{x} \in \mathcal{X}^n, P_{\boldsymbol{x}}^{(2)} = P_{XY}, x_1 = a\}$.

The analog of (II.1) for a Markov chain $X_1, X_2, \cdots$ with stationary transition probabilities given by a matrix W is that if $P_{\boldsymbol{x}}^{(2)} = P_{XY}$ and $x_1 = a$ (with $\Pr\{X_1 = a\} > 0$) then

$$\begin{aligned}\Pr\{X^n = \boldsymbol{x} | X_1 = a\} &= \prod_{(a,b)\in\mathcal{X}^2} W(b|a)^{(n-1)P_{XY}(a,b)} \\ &= \exp\{-(n-1)[H(Y|X) + D(P_{XY}\|W)]\} \end{aligned} \tag{VII.2}$$

where $X^n = X_1 \cdots X_n$. The analog of Lemma II.2 is that for $P_{XY} \in \mathcal{P}_n^{(2)}(\mathcal{X}, a)$

$$|\mathcal{T}_{XY,a}^{n,2}| \approx \exp\{nH(Y|X)\} \tag{VII.3}$$

$$\Pr\{X^n \in \mathcal{T}_{XY,a}^{n,2} | X_1 = a\} \approx \exp\{-nD(P_{XY}\|W)\}. \tag{VII.4}$$

Of course, (VII.4) is a consequence of (VII.2) and (VII.3). The simple idea in the proof of Lemma II.2 suffices only for the $\lesssim$ part of (VII.3), the $\gtrsim$ part is more delicate. One way to get it (Boza [19]) is via the exact formula for $|\mathcal{T}_{XY,a}^{n,2}|$ due to Whittle [80], an elementary proof of which has been given by Billingsley [14]. An important property of second-order types is that they have (equal or) asymptotically equal marginals as $n \to \infty$. Indeed, for $\boldsymbol{x} = ax_2 \cdots x_{n-1}b \in \mathcal{T}_{XY,a}^{n,2}$ the marginals P_X and P_Y of $P_{XY} = P_{\boldsymbol{x}}^{(2)}$ differ only at a and b, if $a \neq b$, both differences being $(n-1)^{-1}$. Moreover, denoting by $\mathcal{P}_e(\mathcal{X}^2)$ the set of those $P \in \mathcal{P}(\mathcal{X}^2)$ whose two marginals are equal, each irreducible $P \in \mathcal{P}_e(\mathcal{X}^2)$ can be arbitrarily approximated by second-order types $P' \in \mathcal{P}_n^{(2)}(\mathcal{X}, a)$ with $P' \ll P$ if n is sufficiently large ($P \in \mathcal{P}_e(\mathcal{X}^2)$ is called irreducible if the stationary Markov chain with two-dimensional distribution P is irreducible).

The above facts permit extensions of the results in Section III to Markov chains, cf. Boza [19], Davisson, Longo, and Sgarro [38], Natarajan [68], Csiszár, Cover, and Choi [26]. In particular, the following analog of Theorem III.3 and its Corollary holds, cf. [26].

Theorem VII.1: Given a Markov chain $X_1, X_2, \cdots$ with transition probability matrix W and $\Pr\{X_1 = a\} > 0$, and a set of PD's $\Pi \subset \mathcal{P}(\mathcal{X}^2)$ such that $P \ll W$ for each $P \in \Pi$, the second-order type $\tilde{P}_n^{(2)}$ of $X_1, \cdots, X_n$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\hat{P}_n \in \Pi | X_1 = a\} = -\min_{P \in \overline{\Pi} \cap \mathcal{P}_e(\mathcal{X}^2)} D(P\|W) \tag{VII.5}$$

iff there exist second-order types $P_n \in \Pi \cap \mathcal{P}_n^{(2)}(\mathcal{X}, a)$ such that $D(P_n\|W)$ approaches the minimum in (VII.5). Further, if the minimum in (VII.5) is attained for a unique P^*, and $X_1^*, X_2^*, \cdots$ denotes a stationary Markov chain with $P_{X_1^* X_2^*} = P^*$, for $(a_1, \cdots, a_k) \in \mathcal{X}^k$ with $P_{X_1^*}(a_1) > 0$, we have

$$\begin{aligned}\lim_{n\to\infty} \Pr\{X_2 = a_2, \cdots, X_k = a_k | \tilde{P}_n^{(2)} \in \Pi, X_1 = a_1\} \\ = \Pr\{X_2^* = a_2, \cdots, X_k^* = a_k | X_1^* = a_1\} \end{aligned} \tag{VII.6}$$

whenever (VII.5) holds for $a = a_1$.

Remarks:

i) Let Π^i denote the set of those irreducible $P \in \Pi \cap \mathcal{P}_e(\mathcal{X}^2)$ for which all $P' \ll P$ in a sufficiently small neighborhood of P belong to Π. The first assertion of Theorem VII.1 gives that (VII.5) always holds if the closure of Π^i equals $\overline{\Pi} \cap \mathcal{P}_e(\mathcal{X}^2)$.

ii) Theorem VII.1 is of interest even if $X_1, X_2 \cdots$ are i.i.d.; the limiting conditional distribution in (VII.6) is Markov rather than i.i.d. also in that case.

As an immediate extension of (VII.1), the rth-order type of a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is defined as the PD $P_{\boldsymbol{x}}^{(r)} \in \mathcal{P}(\mathcal{X}^r)$ with

$$\begin{aligned}&P_{\boldsymbol{x}}^{(r)}(a_1, \cdots, a_r) \\ &\quad = \frac{1}{n-r+1} |\{i\colon x_i = a_1, \cdots, x_{i+r-1} = a_r\}|. \end{aligned} \tag{VII.7}$$

This is the suitable type concept for order-$(r-1)$ Markov chains, in which conditional distributions given the past depend on the last $(r-1)$ symbols. All results about Markov chains and second-order types have immediate extensions to order-$(r-1)$ Markov chains and rth-order types.

Since order-k Markov chains are also order-ℓ ones if $\ell > k$, the analog of the hypothesis-testing result Theorem III.2 can be applied to test the hypothesis that a process known to be Markov of order k_0 is actually Markov of order k for a given $k < k_0$. Performing a multiple test (for each $k < k_0$) amounts to estimating the Markov order. A recent paper analyzing this approach to Markov order estimation is Finesso, Liu, and Narayan [45], cf. also prior works of Gutman [51] and Merhav, Gutman, and Ziv [66].

"Circular" versions of second- and higher order types are also often used as in [38]. The rth-order circular type

of $x_1 \cdots x_n \in \mathcal{X}^n$ is the same as the rth-order type of $x_1 \cdots x_n x_1 \cdots x_{r-1} \in \mathcal{X}^{n+r-1}$, i.e., the joint type of the r sequences $x_1 \cdots x_n$, $x_2 \cdots x_n x_1, \cdots, x_r \cdots x_n \cdots x_{r-1}$. A technical advantage of circular types is compatibility: lower order circular types are marginals of the higher order ones. The price is that expressing probabilities in terms of circular types is more awkward.

B. Finite-State Types

A sequence of $\mathcal{X}$-valued RV's $X_1, X_2, \cdots$ is called a unifilar finite-state source if there exists a (finite) set $\mathcal{S}$ of "states," an initial state $s_1 \in \mathcal{S}$, and a mapping $f\colon \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ that specifies the next state as a function of the present state and source output, such that

$$\Pr\{X^n = \boldsymbol{x}\} = \prod_{i=1}^{n} W(x_i|s_i), \quad s_{i+1} = f(s_i, x_i) \quad \text{(VII.8)}$$

where W is a stochastic matrix specifying the source output probabilities given the states. As the state sequence $\boldsymbol{s} = s_1 \cdots s_n$ is uniquely determined by $\boldsymbol{x} = x_1 \cdots x_n$ and the initial state s_1, so is the joint type $P_{\boldsymbol{sx}}$. It will be called the finite state type $P_{\boldsymbol{x}}^{f, s_1}$ of $\boldsymbol{x}$, given the mapping f and the initial state s_1, cf. Weinberger, Merhav, and Feder [79]. Notice that the rth-order type (VII.7) of $\boldsymbol{x} = x_1 \cdots x_n$ is equal to the finite state type of $x_r \cdots x_n$ for $\mathcal{S} = \mathcal{X}^{r-1}$ and $f\colon \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ defined by $f(a_1 \cdots a_{r-1}, a) = a_2 \cdots a_{r-1} a$, with $s_1 = x_1 \cdots x_{r-1}$.

Denote the set of finite-state types $P_{\boldsymbol{x}}^{f, s_1}$, $\boldsymbol{x} \in \mathcal{X}^n$, by $\mathcal{P}_n^f(\mathcal{X}, s_1)$, and let $\mathcal{T}_{SX, s_1}^{n, f}$ denote the class of sequences $\boldsymbol{x} \in \mathcal{X}^n$ with $P_{\boldsymbol{x}}^{f, s_1} = P_{SX}$. Then for $\boldsymbol{x} \in \mathcal{T}_{SX, s_1}^{n, f}$, (VII.8) gives

$$\Pr\{X^n = \boldsymbol{x}\} = \exp\{-n[H(X|S) + D(P_{SX}\|W)]\}. \quad \text{(VII.9)}$$

Further, for $P_{SX} \in \mathcal{P}_n^f(\mathcal{X}, s_1)$ the following analogs of (VII.3) and (VII.4) hold:

$$|\mathcal{T}_{SX, s_1}^{n, f}| \approx \exp\{nH(X|S)\} \quad \text{(VII.10)}$$

$$\Pr\{X^n \in \mathcal{T}_{SX, s_1}^{n, f}\} \approx \exp\{-nD(P_{SX}\|W)\}. \quad \text{(VII.11)}$$

These permit extensions of results about Markov chains and second-order types to unifilar finite-state sources and finite-state types. Weinberger, Merhav, and Feder [79] used this type concept to study the performance of universal sequential codes for individual sequences (rather than in the averaged sense). They established a lower bound to the codelength, valid for most sequences in any given type class $\mathcal{T}_{SX, s_1}^{n, f}$, except for a vanishingly small fraction of the finite-state types $P_{SX} \in \mathcal{P}_n^f(\mathcal{X}, s_1)$.

The finite-state model (VII.8) can be extended in various ways. Let us consider here the extension when the "next state" s_{i+1} depends on the past sequence $x^i = x_1 \cdots x_i$ not necessarily through s_i and x_i but, more generally, $s_{i+1} = F(x^i)$ where $F\colon \mathcal{X}^* \to \mathcal{S}$ is an arbitrary mapping. Here $\mathcal{X}^*$ denotes the set of all finite sequences of symbols from $\mathcal{X}$, including the void sequence λ; the initial state s_1 need not be explicitly specified in this model, as it is formally given by $s_1 = F(\lambda)$. The type concept adequate for the source model

$$\Pr\{X^n = \boldsymbol{x}\} = \prod_{i=1}^{n} W(x_i|s_i), \qquad s_{i+1} = F(x^i) \quad \text{(VII.12)}$$

is the F-type $P_{\boldsymbol{x}}^F$ defined as the joint type $P_{\boldsymbol{sx}}$, where $\boldsymbol{s}$ is determined by $\boldsymbol{x}$ as in (VII.12). Of course, for the corresponding F-type classes $\mathcal{T}_{SX}^{n, F}$ we still have (VII.9) when $\boldsymbol{x} \in \mathcal{T}_{SX}^{n, F}$, and consequently also

$$|\mathcal{T}_{SX}^{n, F}| \le \exp\{nH(X|S)\}. \quad \text{(VII.13)}$$

Unlike for the finite-state type classes $\mathcal{T}_{SX, s_1}^{n, f}$, however, a lower bound counterpart of (VII.13) cannot be established in general.

An early appearance of this F-type concept, though not of the term, was in Csiszár and Körner [31], applied to DMC's with feedback. The encoder of a feedback code of blocklength n for N messages is defined by mappings $F_k = \mathcal{Y}^* \to \mathcal{X}$, $k = 1, \cdots, N$, that specify the input symbols $x_i = F_k(y^{i-1})$, $i = 1, \cdots, n$, depending on the previous received symbols $y^{i-1} = y_1 \cdots y_{i-1}$, when message k is to be transmitted. Then the received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ in generated by a generalized finite-state model as in (VII.12), with alphabet $\mathcal{Y}$, state set $\mathcal{X}$, and $F = F_k$. In particular, the probability of receiving an $\boldsymbol{y} \in \mathcal{T}_{XY}^{n, F_k}$ equals $\exp\{-n[H(Y|X) + D(P_{XY}\|W)]\}$, cf. (VII.9). Hence a decoder φ will correctly decode message k with probability

$$\sum_{P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})} |\mathcal{T}_{XY}^{n, F_k} \cap D_k| \exp\{-n[H(Y|X) + D(P_{XY}\|W)]\} \quad \text{(VII.14)}$$

where $D_k = \{\boldsymbol{y}\colon \varphi(\boldsymbol{y}) = k\}$. Similarly to (IV.16), we have

$$\sum_{k=1}^{N} |\mathcal{T}_{XY}^{n, F_k} \cap D_k| \le \left| \bigcup_{k=1}^{N} \mathcal{T}_{XY}^{n, F_k} \right| \le |\mathcal{T}_Y^n| \le \exp\{nH(Y)\}. \quad \text{(VII.15)}$$

On account of (VII.13) (with (S, X) replaced by (X, Y)), the left-hand side of (VII.15) is also $\le N \exp\{nH(Y|X)\}$. It follows that if $N > \exp(nR)$ then

$$\frac{1}{N} \sum_{k=1}^{N} |\mathcal{T}_{XY}^{n, F_k} \cap D_k| \le \exp\{n[H(Y|X) - |R - I(X \wedge Y)|^+]\}. \quad \text{(VII.16)}$$

Averaging the probability of correct decoding (VII.14) over the messages $1 \le k \le N$, (VII.16) implies that the average probability of correct decoding is

$$\lesssim \exp\{-n \min[D(P_{XY}\|W) + |R - I(X \wedge Y)|^+]\}$$

where the minimum is for all $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Comparing this with Remark iii) to Theorem IV.2 shows that feedback cannot exponentially improve the probability of correct decoding at rates above channel capacity.

A recent combinatorial result of Ahlswede, Yang, and Zhang [11] is also easiest to state in terms of F-types. Their "inherently typical subset lemma" says, effectively, that given $\mathcal{X}$ and $\varepsilon > 0$, there is a finite set $\mathcal{S}$ such that for sufficiently

large n, to any $A \subset \mathcal{X}^n$ there exists a mapping F: $\mathcal{X}^* \to \mathcal{S}$ and an F-type P_{SX} such that

$$|A| \approx |A \cap \mathcal{T}_{SX}^{n,F}| \geq \exp\{n(H(X|S) - \varepsilon)\}. \quad \text{(VII.17)}$$

While this lemma is used in [11] to prove (the converse part of) a probabilistic result, it is claimed to also yield the asymptotic solution of the general isoperimetric problem for arbitrary finite alphabets and arbitrary distortion measures.

C. Continuous Alphabets

Extensions of the type concept to continuous alphabets are not known. Still, there are several continuous-alphabet problems whose simplest (or the only) available solution relies upon the method of types, via discrete approximations. For example, the capacity subject to a state constraint of an AVC with general alphabets and states, for deterministic codes and the average probability of error criterion, has been determined in this way, cf. [25].

At present, this approach seems necessary even for the following intuitive result.

Theorem VII.2 ([25]): Consider an AVC whose permissible n-length inputs $\boldsymbol{x} \in \mathbb{R}^n$ satisfy $||\boldsymbol{x}||^2 \leq n\Gamma$, and the output is $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z} + Z^n$ where the deterministic sequence $\boldsymbol{z} \in \mathbb{R}^n$ and the random sequence Z^n with independent zero-mean components Z_i may be arbitrary subject to the power constraints $||\boldsymbol{z}||^2 \leq n\Lambda_1$, $\sum_{i=1}^n E(Z_i^2) \leq n\Lambda_2$. This AVC has the same a-capacity as the Gaussian one where the Z_i's are i.i.d. Gaussian RV's with variance Λ_2.

For the latter Gaussian case, Csiszár and Narayan [36] had previously shown that

$$C_a = \begin{cases} \frac{1}{2}\log\left(1 + \frac{\Gamma}{\Lambda_1 + \Lambda_2}\right), & \text{if } \Gamma > \Lambda_1 \\ 0, & \text{if } \Gamma \leq \Lambda_1. \end{cases} \quad \text{(VII.18)}$$

Discrete approximations combined with the method of types provide the simplest available proof of a general form of Sanov's theorem, for RV's with values in an arbitrary set $\mathcal{X}$ endowed with a σ-algebra $\mathcal{F}$ (the discrete case has been discussed in Section III).

For probability measures (pm's) P, Q on $(\mathcal{X}, \mathcal{F})$, the I-divergence $D(P||Q)$ is defined as

$$D(P||Q) = \sup_{\mathcal{A}} D(P^{\mathcal{A}}||Q^{\mathcal{A}}) \quad \text{(VII.19)}$$

the supremum taken for partitions $\mathcal{A} = (A_1, \cdots, A_k)$ of $\mathcal{X}$ into sets $A_i \in \mathcal{F}$. Here $P^{\mathcal{A}}$ denotes the $\mathcal{A}$-quantization of P defined as the distribution $\{P(A_1), \cdots, P(A_k)\}$ on the finite set $\{1, \cdots, k\}$.

The τ-topology of pm's on $(\mathcal{X}, \mathcal{F})$ is the topology in which a pm P belongs to the interior of a set Π of pm's iff for some partition $\mathcal{A} = (A_1, \cdots, A_k)$ and $\varepsilon > 0$

$$\{P'\colon |P'(A_j) - P(A_j)| < \varepsilon, \qquad j = 1, \cdots, k\} \subset \Pi. \quad \text{(VII.20)}$$

The empirical distribution $\hat{P}_n$ of an n-tuple $X^n = (X_1, \cdots, X_n)$ of $\mathcal{X}$-valued RV's is the random pm defined by

$$\hat{P}_n(A) = \frac{1}{n}|\{i\colon X_i \in A\}|, \qquad A \in \mathcal{F}. \quad \text{(VII.21)}$$

Theorem VII.3: Let $X_1, X_2, \cdots$ be independent $\mathcal{X}$-valued RV's with common distribution Q. Then

$$\liminf_{n\to\infty} \frac{1}{n}\log \Pr\{\hat{P}_n \in \Pi\} \geq -\inf_{P \in \Pi^0} D(P||Q), \quad \text{(VII.22)}$$

$$\limsup_{n\to\infty} \frac{1}{n}\log \Pr\{\hat{P}_n \in \Pi\} \leq -\inf_{P \in \overline{\Pi}} D(P||Q), \quad \text{(VII.23)}$$

for every set Π of pm's on $(\mathcal{X}, \mathcal{F})$ for which the probabilities $\Pr\{\hat{P}_n \in \Pi\}$ are defined. Here Π^0 and $\overline{\Pi}$ denote the interior and closure of Π in the τ-topology.

Theorem VII.3 is a general version of Sanov's theorem. In the parlance of large derivations theory (cf. Dembo and Zeitouni [39]) it says that $\{\hat{P}_n\}$ satisfies the large deviation principle with good rate function $D(\cdot||Q)$ ("goodness" means that the sets $\{P\colon D(P||Q) \leq \alpha\}$ are compact in the τ-topology; the easy proof of this property is omitted).

Proof: (Groeneboom, Oosterhoff, and Ruymgaart [49]) Pick any $P \in \Pi^0$, and $\mathcal{A}$ and ε satisfying (VII.20). Apply Theorem III.3 to the quantized RV's $X_i^{\mathcal{A}}$ with distribution $Q^{\mathcal{A}}$, where $X_i^{\mathcal{A}} = j$ if $X_i \in A_j$, and to the set of those distributions $\tilde{P}$ on $\{1, \cdots, k\}$ for which $|\tilde{P}(j) - P(A_j)| < \varepsilon$, $j = 1, \cdots, k$.

As the latter is an open set containing $P^{\mathcal{A}}$, it follows that

$$\lim_{n\to\infty} \frac{1}{n}\log \Pr\{|\hat{P}_n(A_j) - P(A_j)| < \varepsilon,\ j = 1, \cdots, k\} \geq -D(P^{\mathcal{A}}||Q^{\mathcal{A}}). \quad \text{(VII.24)}$$

The left hand side of (VII.24) is a lower bound to that of (VII.22), by (VII.20). Hence, as $P \in \Pi^0$ has been arbitrary, (VII.19) and (VII.24) imply (VII.22).

Notice next that for each partition $\mathcal{A}$, Theorem III.3 applied to the quantized RV's $X_i^{\mathcal{A}}$ as above and to $\Pi^{\mathcal{A}} = \{P^{\mathcal{A}}\colon P \in \Pi\}$ gives that

$$\limsup_{n\to\infty} \frac{1}{n}\log \Pr\{\hat{P}_n^{\mathcal{A}} \in \Pi^{\mathcal{A}}\} \leq -D(\Pi^{\mathcal{A}}||Q^{\mathcal{A}}) = -\inf_{P \in \Pi} D(P^{\mathcal{A}}||Q^{\mathcal{A}}). \quad \text{(VII.25)}$$

Clearly, (VII.23) follows from (VII.19) and (VII.25) if one shows that

$$\sup_{\mathcal{A}} \inf_{P \in \Pi} D(P^{\mathcal{A}}||Q^{\mathcal{A}}) = \inf_{P \in \overline{\Pi}} \sup_{\mathcal{A}} D(P^{\mathcal{A}}||Q^{\mathcal{A}}). \quad \text{(VII.26)}$$

The nontrivial but not too hard proof of (VII.26) is omitted.

The "discrete approximation plus method of types" approach works also for other problems that can not be entered here. For extensions of the hypothesis testing results in Section III, cf. Tusnády [78].

VIII. Conclusions

The method of types has been shown to be a powerful tool of the information theory of discrete memoryless systems. It affords extensions also to certain models with memory, and can be applied to continuous alphabet models via discrete approximations. The close links of the method of types to large deviations theory (primarily to Sanov's theorem) have also been established.

Sometimes it is claimed that "type" arguments, at least for models involving only one set of sequences as in hypothesis testing, could be replaced by referring to general results from large deviations theory. This is true for some applications (although the method of types gives more insight), but in other applications the explicit "type" bounds valid for all n afford stronger conclusions than the asymptotic bounds provided by large deviations theory. It is interesting to note in this respect that for the derivation of (VII.6) even the familiar type bound was not sufficient, rather, the exact formula (of Whittle [80]) for the size of second order type classes had to be used.

Of course, the heavy machinery of large deviations theory (cf. [39]) works for many problems for which type arguments do not. In particular, while that machinery is not needed for Sanov's theorem (Theorem VII.3), it appears necessary to derive the corresponding result for continuous alphabet Markov chains. Indeed, although the method of types does work for finite alphabet Markov chains (cf. Theorem VII.1), extension to general alphabets via discrete approximations does not seem feasible, since quantization destroys the Markov property.

APPENDIX

Proof of Lemma IV.1: Pick $2N$ sequences $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2N}$ from $\mathcal{T}_P^n$ at random. Then, using Lemma II.2, we have for any joint type $P_{X\tilde{X}} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X})$ with $P_X = P_{\tilde{X}} = P$, and any $i \neq j$,

$$\begin{aligned}\Pr\{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}_{X\tilde{X}}^n\} &= \frac{|\mathcal{T}_{X\tilde{X}}^n|}{|\mathcal{T}_P^n|^2} \\ &\leq \frac{\exp\{nH(X,\tilde{X})\}}{[|\mathcal{P}_n(\mathcal{X})|^{-1}\exp\{nH(P)\}]^2} \\ &= |\mathcal{P}_n(\mathcal{X})|^2 \exp\{-nI(X \wedge \tilde{X})\}. \quad \text{(A.1)}\end{aligned}$$

This implies that

$$\begin{aligned}&E|\{j\colon j \neq i, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}_{X\tilde{X}}^n\}| \\ &\quad \leq 2N|\mathcal{P}_n(\mathcal{X})|^2 \exp\{-nI(X \wedge \tilde{X})\}. \quad \text{(A.2)}\end{aligned}$$

Writing

$$\begin{aligned}&F_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2N}) \\ &\quad = \sum_{\substack{P_{X\tilde{X}} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{X}) \\ P_X = P_{\tilde{X}} = P}} |\{j\colon j \neq i, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}_{X\tilde{X}}^n\}| \\ &\qquad \cdot \exp\{nI(X \wedge \tilde{X})\}, \quad \text{(A.3)}\end{aligned}$$

it follows from (A.2) that

$$E\sum_{i=1}^{2N} F_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2N}) \leq 4N^2|\mathcal{P}_n(\mathcal{X})|^2|\mathcal{P}_n(\mathcal{X}\times\mathcal{X})|. \quad \text{(A.4)}$$

On account of (A.4), the same inequality must hold without the expectation sign for some choice of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2N}$, and then the latter satisfy

$$F_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2N}) \leq 4N|\mathcal{P}_n(\mathcal{X})|^2|\mathcal{P}_n(\mathcal{X}\times\mathcal{X})| \quad \text{(A.5)}$$

for at least N indices i. Assuming without any loss of generality that (A.5) holds for $i = 1, \cdots, N$, it follows by (A.3) that

$$\begin{aligned}&|\{j\colon j \neq i, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{T}_{X\tilde{X}}^n\}| \\ &\quad \leq 4N|\mathcal{P}_n(\mathcal{X})|^2|\mathcal{P}_n(\mathcal{X}\times\mathcal{X})| \exp\{-nI(X \wedge \tilde{X})\} \quad \text{(A.6)}\end{aligned}$$

for each $1 \leq i \leq N$ and $P_{X\tilde{X}} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X})$ with $P_X = P_{\tilde{X}} = P$. As $N \approx \exp(nR)$ may be chosen such that $4N|\mathcal{P}_n(\mathcal{X})|^2|\mathcal{P}_n(\mathcal{X}\times\mathcal{X})| < \exp(nR)$, this completes the proof.

Proof of Theorems V.1, V.2 (continued): Consider a decoder as defined in Section V, with a preliminarily unspecified permissible set $\Psi \subset \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$. Recall that Π denotes $\Pi^m(\eta)$ or $\Pi^a(\eta)$ defined by (V.14) and (V.15), according as the maximum or average probability of error criterion is considered, and each $P_{X\tilde{X}SY} \in \Psi$ has to satisfy (V.20).

Clearly, for $\boldsymbol{y} \in \mathcal{T}_{Y|XS}^n(\boldsymbol{x}_i, \boldsymbol{s})$ with $P_{XSY} \in \Pi$ we can have $\varphi(\boldsymbol{y}) \neq i$ only if $\boldsymbol{y} \in \mathcal{T}_{Y|X\tilde{X}S}^n(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s})$ for some $j \neq i$ and $P_{X\tilde{X}SY} \in \Psi \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$. Using (II.7) and (V.10), it follows that the fraction of sequences in $\mathcal{T}_{Y|XS}^n(\boldsymbol{x}_i, \boldsymbol{s})$ with $\varphi(\boldsymbol{y}) \neq i$ is bounded, in the $\lesssim$ sense, by

$$\begin{aligned}&\exp\{-nH(Y|XS)\}\sum \exp\{nH(Y|X\tilde{X}S)\} \\ &\qquad \cdot \exp\{n[|R - I(XS \wedge \tilde{X})|^+ + \varepsilon]\} \\ &\quad \approx \max \exp\{-nI(\tilde{X} \wedge Y|XS) + n|R - I(XS \wedge \tilde{X})|^+ \\ &\qquad + n\varepsilon\}\end{aligned}$$

where the sum and max are for all joint types $P_{X\tilde{X}SY} \in \Psi \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$ with the given marginal P_{XSY}. Except for an exponentially small fraction of the indices $1 \leq i \leq N$, it suffices to take the above sum and max for those joint types that satisfy the additional constraint

$$I(X \wedge \tilde{X}S) \leq |R - I(\tilde{X} \wedge S)|^+ + \varepsilon. \quad \text{(A.7)}$$

Indeed, the fraction of indices $1 \leq i \leq N$ to which a $j \neq i$ exists with $P_{\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{s}} = P_{X\tilde{X}S}$ not satisfying (A.7), is exponentially small by (V.12).

If the fraction of incorrectly decoded $\boldsymbol{y}$'s within $\mathcal{T}_{Y|XS}^n(\boldsymbol{x}_i, \boldsymbol{s})$ is exponentially small for each $P_{XSY} \in \Pi^m(\eta) \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$, it follows by (V.3) and (V.17) that $e_i(\boldsymbol{s})$ is exponentially small. Hence, writing

$$F(X, \tilde{X}, S, Y) = I(\tilde{X} \wedge Y|XS) - |R - I(XS \wedge \tilde{X})|^+, \quad \text{(A.8)}$$

the maximum probability of error e defined by (V.2) will be exponentially small if a $\xi > \varepsilon$ exists such that $F(X, \tilde{X}, S, Y) \geq \xi$ whenever $P_{X\tilde{X}SY} \in \Psi$.

Similarly, if the fraction of incorrectly decoded $\boldsymbol{y}$'s within $\mathcal{T}_{Y|XS}^n(\boldsymbol{x}_i, \boldsymbol{s})$ is exponentially small for each $P_{XSY} \in \Pi^a(\eta) \cap \mathcal{P}_n(\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$, except perhaps for an exponentially small fraction of the indices $1 \leq i \leq N$, it follows by (V.18) that $N^{-1}\sum_i e_i(\boldsymbol{s})$ is exponentially small, supposing, of course, that $\varepsilon < \eta$. Hence the average probability of error $\bar{e}$ defined by (V.2) will be exponentially small if a $\xi > \varepsilon$ exists such that $F(X, \tilde{X}, S, Y) \geq \xi$ whenever (A.7) holds and $P_{X\tilde{X}SY} \in \Psi$.

Actually, in both cases $F(X, \tilde{X}, S, Y) \geq \xi > 0$ suffices, as $\varepsilon > 0$ in Lemma V.1 can always be chosen smaller than ξ.

To complete the proof, it suffices to find a permissible set of joint distributions $\Psi \subset \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y})$ such that

i) in case of Theorem V.1, $F(X, \tilde{X}, S, Y)$ has a positive lower bound subject to $P_{X\tilde{X}SY} \in \Psi$ if $R < \min[C(P), D(P)]$,

ii) in the case of Theorem V.2, $F(X, \tilde{X}, S, Y)$ has a positive lower bound subject to $P_{X\tilde{X}SY} \in \Psi$ and (A.7) if

$$R < \min_{P_{\tilde{S}} \in \mathcal{P}(\mathcal{S})} I(X \wedge Y) \text{ where } P_{XSY} = P \times P_{\tilde{S}} \times W, \tag{A.9}$$

cf. (V.4), and the AVC is nonsymmetrizable.

Now, from (A.8),

$$\begin{aligned} F(X, \tilde{X}, S, Y) &= I(\tilde{X} \wedge YXS) - R \\ &\geq I(\tilde{X} \wedge Y) - R \quad \text{if } I(XS \wedge \tilde{X}) \leq R. \end{aligned} \tag{A.10}$$

Moreover, if (A.7) holds then $I(\tilde{X} \wedge S) \leq R$ implies $R + \varepsilon \geq I(X \wedge \tilde{X}S) + I(\tilde{X} \wedge S) = I(\tilde{X} \wedge XS) + I(X \wedge S) \geq I(\tilde{X} \wedge XS)$, hence

$$\begin{aligned} F(X, \tilde{X}, S, Y) &\geq I(\tilde{X} \wedge YXS) - R - \varepsilon \\ &\geq I(\tilde{X} \wedge Y) - R - \varepsilon \quad \text{if } I(\tilde{X} \wedge S) \leq R. \end{aligned} \tag{A.11}$$

If $P_{X\tilde{X}SY} \in \Psi$ then $P_{\tilde{X}Y}$ is the marginal of some $P_{\tilde{X}\tilde{S}Y} \in \Pi$, cf. (V.20), where Π denotes $\Pi^m(\eta)$ or $\Pi^a(\eta)$. In the first case, $P_{\tilde{X}\tilde{S}Y} \in \Pi = \Pi^m(\eta)$ implies by (V.14) that $P_{\tilde{X}\tilde{S}Y}$ is arbitrarily close to $P_{\tilde{X}\tilde{S}} \times W$ and hence any number less than $C(P)$ defined by (V.7) is a lower bound to $I(\tilde{X} \wedge Y)$ if η is sufficiently small. Then (A.10) shows that the claim under i) always holds when $I(XS \wedge \tilde{X}) \leq R$. In the second case $P_{\tilde{X}\tilde{S}Y} \in \Pi = \Pi^a(\eta)$ implies by (V.15) that $P_{\tilde{X}\tilde{S}Y}$ is close to $P \times P_{\tilde{S}_s} \times W$ and hence any number less than the minimum in (A.9) is a lower bound to $I(\tilde{X} \wedge Y)$ if η is sufficiently small. Then (A.11) shows that the claim under ii) always holds when $I(\tilde{X} \wedge S) \leq R$.

So far, the choice of Ψ played no role. To make the claim under i) hold also when $I(XS \wedge \tilde{X}) > R$, chose Ψ as the set of joint distributions $P_{X\tilde{X}SY}$ satisfying $I(\tilde{X} \wedge Y|XS) \geq \xi > 0$, in addition to (V.20). It can be shown by rather straightforward calculation using (V.8) and (V.13) that this Ψ is permissible if $R < D(P)$, providing η and ξ are sufficiently small, cf. [29] for details.

Concerning the claim under ii) in the remaining case $I(\tilde{X} \wedge S) > R$, notice that then (A.8) gives

$$\begin{aligned} F(X, \tilde{X}, S, Y) &= I(\tilde{X} \wedge Y|XS) \\ &= I(\tilde{X} \wedge YX|S) - I(\tilde{X} \wedge X|S) \\ &\geq I(\tilde{X} \wedge YX|S) - \varepsilon \end{aligned}$$

because $I(\tilde{X} \wedge X|S) \leq I(X \wedge \tilde{X}S) \leq \varepsilon$ by (A.7). Hence the claim will hold if Ψ is chosen as the set of those joint distributions $P_{X\tilde{X}SY}$ that satisfy $I(\tilde{X} \wedge YX|S) \geq \xi > 0$ in addition to (V.20). It can be shown that this Ψ is permissible if the AVC is nonsymmetrizable and η, ξ are sufficiently small, cf. [33] for details.

REFERENCES

[1] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Inform. Theory* (Tsahkadzor, Armenian SSR, 1971). Budapest, Hungary: Akadémiai Kiadó, 1973, pp. 23–52.

[2] ———, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitsth. Verw. Gebiete*, vol. 44, pp. 159–175, 1978.

[3] ———, "Coloring hypergraphs: A new approach to multi-user source coding I–II," *J. Comb. Inform. Syst. Sci.*, vol. 4, pp. 76–115, 1979, and vol. 5, pp. 220–268, 1980.

[4] ———, A method of coding and an application to arbitrarily varying channels, *J. Comb. Inform. Systems Sci.*, vol. 5, pp. 10–35, 1980.

[5] ———, "Extremal properties of rate distortion functions," *IEEE Trans. Inform. Theory*, vol. 36, pp. 166–171, Jan. 1990.

[6] ———, "The maximal error capacity of arbitrarily varying channels for constant list sizes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1416–1417, 1993.

[7] R. Ahlswede and N. Cai, "Two proofs of Pinsker's conjecture concerning arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1647–1649, Nov. 1991.

[8] ———, "Arbitrarily varying multiple-access channels, part 1. Ericson's symmetrizability is adequate, Gubner's conjecture is true," Preprint 96–068, Sonderforschungsbereich 343, Universität Bielefeld, Bielefeld, Germany.

[9] R. Ahlswede, N. Cai, and Z. Zhang, "Erasure, list, and detection zero-error capacities for low noise and a relation to identification," *IEEE Trans. Inform. Theory*, vol. 42, pp. 52–62, Jan. 1996.

[10] R. Ahlswede and J. Wolfowitz, "The capacity of a channel with arbitrarily varying cpf's and binary output alphabets," *Z. Wahrscheinlichkeitsth. Verw. Gebiete*, vol. 15, pp. 186–194, 1970.

[11] R. Ahlswede, E. Yang, and Z. Zhang, "Identification via compressed data," *IEEE Trans. Inform. Theory*, vol. 43, pp. 48–70, Jan. 1997.

[12] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1989–1902, Nov. 1995.

[13] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NJ: Prentice Hall, 1971.

[14] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, pp. 12–40; correction, p. 1343, 1961.

[15] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.

[16] R. E. Blahut, "Composition bounds for channel block codes," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 656–674, 1977.

[17] V. Blinovsky, P. Narayan, and M. Pinsker, "Capacity of the arbitrarily varying channel under list decoding," *Probl. Pered. Inform.*, vol. 31, pp. 99–113, 1995.

[18] L. Boltzmann, "Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht," *Wien. Ber.*, vol. 76, pp. 373–435, 1877.

[19] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *Ann. Math. Statist.*, vol. 42, pp. 1992–2007, 1971.

[20] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[21] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes.* Amsterdam, The Netherlands: North Holland, 1997.

[22] I. Csiszár, "Joint source-channel error exponent," *Probl. Contr. Inform. Theory*, vol. 9, pp. 315–328, 1980.

[23] ———, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 585–592, July 1982.

[24] ———, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 823–828, Nov. 1982.

[25] ———, "Arbitrarily varying channels with general alphabets and states," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1725–1742, 1992.

[26] 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.

[27] I. Csiszár and J. Körner "Towards a general theory of source networks," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 155–165, Jan. 1980.

[28] ———, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–11, Jan. 1981.
[29] ———, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitsth. Verw. Gebiete*, vol. 57, pp. 87–101, 1981.
[30] ———, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[31] ———, "Feedback does not affect the reliability function of a DMC at rates above capacity," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 92–93, Jan. 1982.
[32] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," presented at the IEEE Int. Symp. Information Theory (Cornell Univ., Ithaca, NY, 1977), preprint.
[33] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Positivity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, Mar. 1988.
[34] ———, "Capacity and decoding rules for arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 752–769, July 1989.
[35] ———, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, pp. 35–43, Jan. 1995.
[36] ———, "Capacity of the Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 37, pp. 18–26, Jan. 1991.
[37] I. Csiszár and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2065–2072, Nov. 1996.
[38] L. 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. 431–348, 1981.
[39] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Jones and Bartlett, 1993.
[40] R. L. Dobrushin and S. Z. Stambler, "Coding theorems for classes of arbitrarily varying discrete memoryless channels," *Probl. Pered. Inform.*, vol. 11, no. 2, pp. 3–22, 1975, English translation.
[41] G. Dueck, "Maximal error capacity regions are smaller than average error capacity regions for multi-user channels," *Probl. Contr. Inform. Theory*, vol. 7, pp. 11–19, 1978.
[42] G. Dueck and J. Körner, "Reliability function of a discrete memoryless channel at rates above capacity," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 82–85, Jan. 1979.
[43] T. Ericson, "Exponential error bounds for random codes in the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 42–48, Jan. 1985.
[44] R. M. Fano, *Transmission of Information, A Statistical Theory of Communications*. New York: Wiley, 1961.
[45] L. Finesso, C. C. Liu, and P. Narayan, "The optimal error exponent for Markov order estimation," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1488–1497, Sept. 1996.
[46] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.
[47] ———, "A perspective on multiaccess channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 124–142, Mar. 1985.
[48] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.
[49] P. Groeneboom, J. Oosterhoff, and F. H. Ruymgaart, "Large deviation theorems for empirical probability measures," *Ann. Probab.*, vol. 7, pp. 553–586, 1979.
[50] J. A. Gubner, "On the deterministic-code capacity of the multiple-access arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 36, pp. 262–275, Mar. 1990.
[51] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401–408, Mar. 1989.
[52] E. A. Haroutunian, "Bounds on the error probability exponent for semicontinuous memoryless channels," *Probl. Pered. Inform.*, vol. 4, no. 4, pp. 37–48, 1968, in Russian.
[53] W. Hoeffding, "Asymptotically optimal tests for multinominal distributions," *Ann. Math. Statist.*, vol. 36, pp. 1916–1921, 1956.
[54] B. L. Hughes, "The smallest list for the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 43, pp. 803–815, May 1997.
[55] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, MIT, Cambridge, MA, 1983.
[56] J. H. Jahn, "Coding for arbitrarily varying multiuser channels," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 212–226, Mar. 1981.
[57] J. Kiefer and J. Wolfowitz, "Channels with arbitrarily varying channel probability functions," *Inform. Contr.*, vol. 5, pp. 44–54, 1962.
[58] J. Körner and A. Sgarro, "Universally attainable error exponents for broadcast channels with degraded message sets," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 670–679, 1980.
[59] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1439–1452, Sept. 1996.
[60] H. J. Liao, "Multiple access channels," Ph.D. dissertation, Univ. Hawaii, Honolulu, 1972.
[61] T. Linder, G. Lugosi, and K. Zeger, "Fixed-rate universal lossy source coding and rates of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 665–676, May 1995.
[62] Y. S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 376–386, 1996.
[63] G. Longo and A. Sgarro, "The source coding theorem revisited: A combinatorial approach," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 544–548, May 1979.
[64] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 197–199, Mar. 1974.
[65] N. Merhav, "A comment on 'A rate of convergence result for a universal d-semifaithful code,'" *IEEE Trans. Inform. Theory*, vol. 41, pp. 1200–1202, July 1995.
[66] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1014–1019, Sept. 1989.
[67] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.
[68] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 360–365, 1985.
[69] Y. Oohama, "Universal coding for correlated sources with linked encoders," *IEEE Trans. Inform. Theory*, vol. 42, pp. 837–847, May 1996.
[70] Y. Oohama and T. S. Han, "Universal coding for the Slepian–Wolf data compression system and the strong converse theorem," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1908–1919, 1994.
[71] J. Pokorny and H. M. Wallmeier, "Random coding bound and codes produced by permutations for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 741–750, Nov. 1985.
[72] I. N. Sanov, "On the probability of large deviations of random variables," *Mat. Sbornik*, vol. 42, pp. 11–44, 1957, in Russian; English translation in *Select. Transl. Math. Statist. Probab.*, vol. 1, pp. 213–244, 1961.
[73] E. Schrödinger, "Über die Umkehrung der Naturgesetze," *Sitzungsber. Preuss. Akad. Wiss. Berlin Phys. Math. Klass.*, vols. 8/9, pp. 144–153, 1931.
[74] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Inform. Contr.*, vol. 10, pp. 65–104 and 522–552, 1967.
[75] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. Illinois Press, 1949.
[76] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, 1973.
[77] İ. E. Telatar and R. G. Gallager, "New exponential upper bounds to error and erasure probabilities," in *Proc. IEEE Int. Symp. Information Theory* (Trondheim, Norway, 1994), p. 379.
[78] G. Tusnády, "On asymptotically optimal tests," *Ann. Statist.*, vol. 5, pp. 358–393, 1977.
[79] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.
[80] P. Whittle, "Some distributions and moment formulae for the Markov chain," *J. Roy. Stat. Soc.*, ser. B, vol. 17, pp. 235–242, 1955.
[81] J. Wolfowitz, *Coding Theorems of Information Theory*. Berlin, Germany: Springer, 1961.
[82] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, pp. 513–820, May 1993.
[83] Z. Zhang, E. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part one: Known statistics," *IEEE Trans. Inform. Theory*, vol. 43, pp. 71–91, Jan. 1997.