The Capacity of the Arbitrarily Varying Channel Revisited: Positivity, Constraints

IMRE CSISZÁR AND PRAKASH NARAYAN, MEMBER, IEEE

Abstract — **A well-known result of Ahlswede asserts that the deterministic code capacity of an arbitrarily varying channel (AVC), under the average error probability criterion, either equals its random code capacity or else is zero. A necessary and sufficient condition is identified for deciding between these alternatives, namely, the capacity is zero if and only if the AVC is symmetrizable. The capacity of the AVC is also determined with constraints on the transmitted codewords as well as on the channel state sequences, and it is demonstrated that it may be positive but less than the corresponding random code capacity. A special case of the results resolves a weakened version of a fundamental problem of coding theory.**

I. Introduction

ARBITRARILY varying channels (AVC's) were introduced by Blackwell *et al.* [3] to model communication channels with unknown parameters which may vary with time in an arbitrary and unknown manner during the transmission of a codeword. Formally, a (discrete memoryless) AVC is determined by a family $\{W(\cdot|\cdot, s), s \in \mathcal{S}\}$ of channels with (finite) input alphabet $\mathcal{X}$ and (finite) output alphabet $\mathcal{Y}$, the individual channels in this family being identified by an index $s \in \mathcal{S}$ called the *state*. Thus $W(y|x, s)$ is the probability that $y \in \mathcal{Y}$ is received given that $x \in \mathcal{X}$ is transmitted and $s \in \mathcal{S}$ is the state of the channel. We shall assume that the set $\mathcal{S}$ of possible states is also finite. For length-n sequences the probability of receiving $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$, when $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is transmitted and $\boldsymbol{s} = (s_1, \cdots, s_n) \in \mathcal{S}^n$ is the channel state sequence, is defined as

$$W^n(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}) = \prod_{k=1}^{n} W(y_k|x_k, s_k). \tag{1.1}$$

Arbitrarily varying channels afford a wide variety of challenging problems to information theorists. The coding problems for the AVC vary according to the kinds of permissible coding strategies and the nature of the performance criteria. Some of these problems are extremely difficult. For instance, Shannon's famous zero-error problem [11], as observed by Ahlswede [1], is a special case of an AVC-capacity problem. The same is true of the fundamental problem of coding theory concerning the largest possible rate of binary codes capable of correcting a fixed fraction of bit errors, as will be indicated in this paper. For a summary of the work on AVC's and for basic results, we refer the reader to Ahlswede [2], Wolfowitz [12], and Csiszár–Körner [4]. Much of our terminology is adopted from Csiszár–Körner [4].

In a previous paper (Csiszár–Narayan [7]), we investigated the effects of various types of constraints on the transmitted or state sequences on the capacity of an AVC. The code was not permitted to depend on the states (i.e., both the encoder and decoder were completely ignorant of the actual state sequence); however, *random codes* (i.e., correlated randomization in encoding and decoding) were permitted. Here we dispense with the last assumption and determine the capacity of the AVC for *deterministic codes* using, as the performance criterion, the *average probability of error*. In doing so, we consider constraints on individual sequences, for having solved this case, other types of constraints can be treated as in [7]. The capacity considered in this paper is called the a-capacity in [4], as distinct from the capacity for the maximum probability of error performance criterion, called the m-capacity. It is a well-known fact that for an AVC, unlike for a simple (discrete memoryless) channel, these two kinds of capacity may differ. In particular, the a-capacity may be positive when the m-capacity is zero. An example due to Ahlswede [2] is the deterministic AVC with $\mathcal{X} = \mathcal{Y} = \{0,1,2,\}$, $\mathcal{S} = \{0,1\}$, $y = x + s$ modulo 3. For notational convenience we shall use the term "capacity" without further specification as referring to a-capacity.

In the absence of any constraints, a celebrated result of Ahlswede [2] asserts that the capacity of an AVC either equals its random code capacity or else is zero. Unfortunately, as Ericson [9] remarks, many AVC's of practical importance are *symmetric* in the sense that $\mathcal{X} = \mathcal{S}$, and $W(y|x, x') = W(y|x', x)$ for every (x, x', y); such AVC's have capacity equal to zero. Reasonable models of real communication situations can then be obtained by imposing constraints on the sequence of channel states, and this may lead to a positive capacity. In this case, however, the proof technique of Ahlswede [2] does not work. In fact, our results will demonstrate that the capacity under a state constraint *may be positive but less than the corresponding random code capacity.*

Manuscript received December 28, 1986; revised April 15, 1987. This work was supported in part by the Systems Research Center at the University of Maryland under NSF Grant OIR-85-00108 and by the Minta Martin Fund for Aerospace Research from the University of Maryland. This paper was presented in part at the 20th Annual Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, MD, March 1987.

I. Csiszár is with the Mathematical Institute of the Hungarian Academy of Sciences, H-1364 Budapest, POB 127, Hungary.

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

IEEE Log Number 8820326.

We first provide a new proof of the basic capacity theorem for the AVC, which also yields, as a new result, a *necessary and sufficient condition for the capacity to be positive*. Our proof employs the method of types, as developed in Csiszár–Körner [4] (following Csiszár–Körner–Marton [6]). A good codeword set is identified by plain random selection, using the bounding technique of Dobrushin–Stambler [8]. The latter, limited by a suboptimal decoding rule, had determined the capacity of the AVC only under rather restrictive conditions. The main new idea in this paper consists of a more subtle decoding rule similar to that in Csiszár–Körner [5], which enables us to bound the error probability as in [5]. The result easily extends to the case when constraints are imposed on the codeword and state sequences.

Our results are formally stated in Section II and proved in Section III. Readers not interested in the details of proofs are advised to proceed from Section II to Section IV, where some interesting implications of the main results for a few simple cases are discussed. More examples and the Gaussian AVC will be treated elsewhere.

II. Preliminaries and Statement of Main Results

We have adopted our terminology from Csiszár–Körner [4]. In particular, all logarithms and exponentials are taken to the base 2.

The message set of a code is identified as the set $\{1,\cdots,N\}$ of positive integers, so that a length-n block code is given by a family of codewords $x_1,\cdots,x_N$, each $\mathcal{X}^n$, and a decoder ϕ: $\mathcal{Y}^n \to \{0,1,\cdots,N\}$. While zero is allowed as a decoder output for the sake of convenience, it always constitutes an error. The probability of error for message i, when this code is used on an AVC defined by (1.1), and the actual state sequence is given to be $s \in \mathcal{S}^n$, equals

$$e(i,s) = \sum_{y:\,\phi(y)\neq i} W^n(y|x_i,s), \tag{2.1}$$

and the *average probability of error* for a state sequence s is

$$\bar{e}(s) = \frac{1}{N}\sum_{i=1}^{N} e(i,s). \tag{2.2}$$

Definition 1: A number $R>0$ is called an *achievable rate* for the given AVC (for deterministic codes and average probability of error criterion) if for every $\epsilon>0$, $\delta>0$, and sufficiently large n, length-n block codes exist with

$$\frac{1}{n}\log N > R-\delta, \qquad \max_{s\in\mathcal{S}^n} \bar{e}(s) \le \epsilon. \tag{2.3}$$

The maximum achievable rate is called the *capacity* of the AVC and is denoted by C.

For $\eta \ge 0$, we define a family of joint distributions P_{XSY} of random variables X, S, and Y with values in $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{Y}$, respectively, by

$$\mathcal{C}_\eta = \{P_{XSY}\colon D(P_{XSY}\|P_X\times P_S\times W) \le \eta\}. \tag{2.4}$$

Here D denotes (Kullback–Leibler) information divergence, and $P_X\times P_S\times W$ denotes a joint distribution on $\mathcal{X}\times\mathcal{S}\times\mathcal{Y}$ with probability mass function $P_X(x)P_S(s)W(y|x,s)$. In particular, $P_{XSY}\in\mathcal{C}_0$ if and only if

$$P_{XSY}(x,s,y) = P_X(x)P_S(s)W(y|x,s). \tag{2.5}$$

Further, we define, for any distribution P on $\mathcal{X}$, the quantity

$$I(P) = \min_{\substack{Y:\ P_{XSY}\in\mathcal{C}_0 \\ \text{for some } S,\text{ with } P_X=P}} I(X\wedge Y). \tag{2.6}$$

Proposition A (Ahlswede): The capacity of the AVC is either $C=\max_P I(P)$ or else $C=0$.

A necessary and sufficient computable characterization of AVC's with $C=0$ does not appear in the literature. The next theorem fills this hiatus; furthermore, we prove it without relying on Proposition A or on the fact (essentially used in Ahlswede's proof [2]) that $\max_P I(P)$ is the random code capacity of the AVC. Note that $\max_P I(P)>0$ and $C=0$ could well occur. Indeed, $\max_P I(P)>0$ holds for many symmetric AVC's, e.g., for the AVC of Example 2 in Section IV, whereas $C=0$ always for a symmetric AVC.

Definition 2: An AVC is *symmetrizable* if for some channel U: $\mathcal{X}\to\mathcal{S}$,

$$\sum_{s\in\mathcal{S}} W(y|x,s)U(s|x') = \sum_{s\in\mathcal{S}} W(y|x',s)U(s|x), \quad \text{for every } x,x',y. \tag{2.7}$$

Theorem 1: $C>0$ if and only if the AVC is not symmetrizable. If $C>0$, then

$$C = \max_P I(P). \tag{2.8}$$

The terminology of Definition 2 is motivated by the fact that if a new AVC, with the set of states coinciding with the input alphabet, is defined by

$$V(y|x,x') = \sum_{s\in\mathcal{S}} W(y|x,s)U(s|x'),$$

then (2.7) states that this new AVC is symmetric. The necessity of nonsymmetrizability for $C>0$ was observed by Ericson [9]. He also compared this necessary condition with the sufficient condition of Ahlswede [2], namely, that two distributions P_1 and P_2 exist on the input alphabet $\mathcal{X}$ such that for any pair of distributions Q_1, Q_2 on the state space $\mathcal{S}$,

$$\sum_{x,s} P_1(x)Q_1(s)W(y|x,s) \neq \sum_{x,s} P_2(x)Q_2(x)W(y|x,s), \quad \text{for at least one } y\in\mathcal{Y}.$$

Ericson's analysis led to the plausibility of this condition being, in general, strictly stronger than nonsymmetrizability, and therefore a necessary and sufficient condition for $C>0$ could not be established. Note, however, that he did not actually prove Ahlswede's [2] sufficient condition to be stronger than nonsymmetrizability; nor do we. We need not address this question as nonsymmetrizability, a simple

condition whose verification involves only linear equations, is proven to be both necessary and sufficient for $C > 0$.

We observe that if the channel is nonsymmetrizable, then $I(P)$ defined by (2.6) is positive for every P satisfying $P(x) > 0$ for all $x \in \mathcal{X}$. Indeed, if $I(P)$ were zero for such a P, then (2.6) implies the existence of a random variable S such that for P_{XSY} defined by (2.5), X and Y are independent. Thus by (2.5), $\sum_{s \in \mathcal{S}} W(y|x, s)P_S(s) = P_Y(y)$ would not depend on x. However, this implies symmetrizability of the channel in a trivial manner, with $U(\cdot|x) = P_s(\cdot)$, not depending on x, which leads to a contradiction.

We recall from [4] that the *type* of a sequence $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is a distribution $P_{\boldsymbol{x}}$ on $\mathcal{X}$ where $P_{\boldsymbol{x}}(x)$ is the relative frequency of x in $\boldsymbol{x}$. Similarly, *joint types* are distributions on product spaces. For example, the joint type of three given sequences $\boldsymbol{x} \in \mathcal{X}^n$, $\boldsymbol{s} \in \mathcal{S}^n$, $\boldsymbol{y} \in \mathcal{Y}^n$ is a distribution $P_{\boldsymbol{x},\boldsymbol{s},\boldsymbol{y}}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ where $P_{\boldsymbol{x},\boldsymbol{s},\boldsymbol{y}}(x, s, y)$ is the relative frequency of the triple (x, s, y) among the triples (x_i, s_i, y_i), $i = 1, \cdots, n$.

In the proof of Theorem 1 good codes are obtained by randomly selecting codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from the set of sequences of a fixed type; the key part consists of finding a suitable decoder ϕ. We use a decoder ϕ defined as follows.

Definition 3: Given the codewords $\boldsymbol{x}_i$, $i = 1, \cdots, N$, let $\phi(\boldsymbol{y}) = i$ if and only if an $\boldsymbol{s} \in \mathcal{S}^n$ exists such that

1) the joint type $P_{\boldsymbol{x}_i,\boldsymbol{s},\boldsymbol{y}}$ belongs to $\mathcal{C}_\eta$ (cf. (2.4));
2) for each competitor $j \neq i$, i.e., such that $P_{\boldsymbol{x}_j,\boldsymbol{s}',\boldsymbol{y}} \in \mathcal{C}_\eta$ for some $\boldsymbol{s}' \in \mathcal{S}^n$, we have $I(XY \wedge X'|S) \leq \eta$, where X, X', S, Y denote dummy random variables such that the joint type of $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s}, \boldsymbol{y})$ equals $P_{XX'SY}$.

If no such i exists, we set $\phi(\boldsymbol{y}) = 0$ (i.e., declare an error).

A main step of the proof of Theorem 1 will consist in showing that this decoding rule is unambiguous if η is sufficiently small.

We observe that condition 1) is a *joint typicality* condition, that is, we require that $(\boldsymbol{x}_i, \boldsymbol{s}, \boldsymbol{y})$ be jointly typical for some $\boldsymbol{s}$ and for a joint distribution of the form (2.5). Dobrushin–Stambler [8] had employed a decoding rule based on similar joint typicality, but their method of eliminating ambiguities in decoding (by simply adopting the smallest i that satisfied the joint typicality condition) did not lead to definitive results. Our condition 2) is analogous to condition (4.10) in Csiszár–Körner [5], where $I(Y \wedge X'|XS)$ was required to be small; here, we additionally ask that $I(X \wedge X'|S)$ also be small.

Let us now consider AVC's with *input* or *state constraints*. As in Csiszár–Narayan [7], let $g(x)$ and $l(s)$ be given functions on $\mathcal{X}$ and $\mathcal{S}$, respectively. For $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{s} = (s_1, \cdots, s_n)$, we define

$$g(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \tag{2.9}$$

$$l(\boldsymbol{s}) = \frac{1}{n} \sum_{i=1}^{n} l(s_i). \tag{2.10}$$

For convenience, we assume as in [7] that

$$\min_{x \in \mathcal{X}} g(x) = \min_{s \in \mathcal{S}} l(s) = 0. \tag{2.11}$$

Definition 4: A number $R > 0$ is an *achievable rate under input constraint* Γ *and state constraint* Λ if for any $\epsilon > 0$, $\delta > 0$, and sufficiently large n, there exist codes with codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$, each satisfying $g(\boldsymbol{x}_i) \leq \Gamma$, and such that

$$\frac{1}{n} \log N > R - \delta, \qquad \max_{\boldsymbol{s}:\, l(\boldsymbol{s}) \leq \Lambda} \bar{e}(\boldsymbol{s}) \leq \epsilon. \tag{2.12}$$

The largest of such achievable rates is called the *capacity of the AVC under input constraint* Γ *and state constraint* Λ; it is denoted by $C(\Gamma, \Lambda)$.

If $\Gamma \geq g_{\max} = \max_{x \in \mathcal{X}} g(x)$ resp. $\Lambda \geq l_{\max} = \max_{s \in \mathcal{S}} l(s)$, then the input resp. state constraint is inoperative. Thus $C(g_{\max}, \Lambda)$ denotes the capacity with state constraint Λ and no input constraint, while $C(\Gamma, l_{\max})$ denotes the capacity with input constraint Γ and no state constraint.

The capacity of the AVC under state constraint Λ may be positive even for symmetrizable (or symmetric) AVC's. Indeed, for the existence of codes with codewords of type P satisfying (2.12) for some $R > \delta$, the crucial question is whether Λ is larger or smaller than

$$\Lambda_0(P) = \min_{U \in \mathcal{U}} \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} P(x)U(s|x)l(s) \tag{2.13}$$

where $\mathcal{U}$ denotes the set of all channels $U: \mathcal{X} \to \mathcal{S}$ satisfying (2.7). Clearly, $\Lambda_0(P)$ is a continuous function of P if $\mathcal{U} \neq \phi$, i.e., if the AVC is symmetrizable and $\Lambda_0(P) = \infty$ for a nonsymmetrizable AVC. Lemma 1 yields, under state constraint Λ, that no code with codewords of type P satisfying $\Lambda_0(P) < \Lambda$ can be "good." On the other hand, if $\Lambda_0(P) > \Lambda$, Theorem 2 asserts that good codes do exist.

Lemma 1: Any code of block length n with $N \geq 2$ codewords, each of type P, with $\Lambda_0(P) < \Lambda$, has

$$\max_{\boldsymbol{s}:\, l(\boldsymbol{s}) \leq \Lambda} \bar{e}(\boldsymbol{s}) \geq \frac{N-1}{2N} - \frac{1}{n} \frac{l_{\max}^2}{(\Lambda - \Lambda_0(P))^2}.$$

In particular, for any $\epsilon < 1/2$,

$$\max_{\boldsymbol{s}:\, l(\boldsymbol{s}) \leq \Lambda} \bar{e}(\boldsymbol{s}) \geq \epsilon, \qquad \text{if } N \geq \frac{2}{1-2\epsilon},$$

$$n \geq \frac{4l_{\max}^2}{(1-2\epsilon)(\Lambda - \Lambda_0(P))^2}.$$

To describe our main result for state constraint Λ, let us denote the set of joint distributions $P_{XSY} \in \mathcal{C}_\eta$ with $El(S) \leq \Lambda$ by $\mathcal{C}_\eta(\Lambda)$, where $\eta \geq 0$. Then $\mathcal{C}_0(\Lambda)$ is the set of joint distributions as in (2.5) for which $El(S) \leq \Lambda$.

For any distribution P on $\mathcal{X}$, and $\Lambda > 0$, we define

$$I(P, \Lambda) = \min_{\substack{Y:\, P_{XSY} \in \mathcal{C}_0(\Lambda) \\ \text{for some } S, \text{ with } P_X = P}} I(X \wedge Y). \tag{2.14}$$

Lemma 2: For any $\Lambda > 0$, $\delta > 0$, and $\epsilon < 1$, there exists n_0 such that for any code of block length $n \geq n_0$ with N

codewords, each of type P,

$$\frac{1}{n}\log N \geq I(P,\Lambda)+\delta \quad \text{implies} \quad \max_{s:l(s)\leq\Lambda} \bar{e}(s) > \epsilon.$$

The following Theorem 2 is our main technical result. Informally, it asserts that if $\Lambda_0(P) > \Lambda$, then $I(P, \Lambda)$, the largest coding rate allowed by Lemma 2, is indeed achievable under state constraint Λ by codes whose codewords are all of type P. In fact, this holds even with an exponentially decreasing probability of error.

Theorem 2: Given $\Lambda > 0$ and arbitrarily small $\alpha > 0$, $\beta > 0$, $\delta > 0$, for any block length $n \geq n_0$ and for any type P with

$$\Lambda_0(P) \geq \Lambda + \alpha, \qquad \min_{x\in\mathcal{X}} P(x) \geq \beta, \tag{2.15}$$

there exists a code with codewords $x_1, \cdots, x_N$, each of type P, such that

$$\frac{1}{n}\log N > I(P,\Lambda) - \delta,$$

$$\max_{s:\, l(s)\leq\Lambda} \bar{e}(s) \leq \exp(-n\gamma) \tag{2.16}$$

where n_0 and $\gamma > 0$ depend only on α, β, and δ, and the given AVC.

The proof of Theorem 2 is similar to that of Theorem 1, replacing $\mathcal{C}_\eta$ by $\mathcal{C}_\eta(\Lambda)$ in the definition of the decoding rule.

The following result on capacity under input constraint Γ and state constraint Λ is a facile consequence of Theorem 2. For notational convenience, we define

$$g(P) = \sum_{x\in\mathcal{X}} P(x)g(x); \tag{2.17}$$

then we have

$$g(x) = g(P_x) \text{ for every } x \in \mathcal{X}^n. \tag{2.18}$$

Theorem 3: For any $\Gamma > 0$, $\Lambda > 0$,

1) $$C(\Gamma,\Lambda) = 0, \quad \text{if } \max_{P:\, g(P)\leq\Gamma} \Lambda_0(P) < \Lambda;$$

2) $$C(\Gamma,\Lambda) = \max_{\substack{P:\, g(P)\leq\Gamma \\ \Lambda_0(P)\geq\Lambda}} I(P,\Lambda) > 0, \quad \text{if } \max_{P:\, g(P)\leq\Gamma} \Lambda_0(P) > \Lambda.$$

The case when $\max_{P:\, g(P)\leq\Gamma} \Lambda_0(P) = \Lambda$ remains unsolved; it appears likely that $C(\Gamma, \Lambda) = 0$. However, at present this can only be proved for special cases, cf. the remark following the proof of Theorem 3.

For the capacity under state constraint Λ and with no input constraint, i.e., for $C(g_{\max}, \Lambda)$, Theorem 3 yields the following.

Corollary: For any $\Lambda > 0$, with Λ_0 denoting $\max_P \Lambda_0(P)$,

1) $C(g_{\max}, \Lambda) = 0$, if $\Lambda_0 < \Lambda$;

2) $C(g_{\max}, \Lambda) = \max_{P:\, \Lambda_0(P)\geq\Lambda} I(P,\Lambda)$ if $\Lambda_0 > \Lambda$.

We know that the random code capacity of the AVC under input constraint Γ and state constraint Λ is

$$C_r(\Gamma,\Lambda) = \max_{P:\, g(P)\leq\Gamma} I(P,\Lambda) \tag{2.19}$$

(cf. [7, theorem 1] where the random code capacity was denoted by $C(\Gamma, \Lambda)$). In particular, the random code capacity under state constraint Λ and with no input constraint (obtained by setting $\Gamma = g_{\max}$) is $\max_P I(P, \Lambda)$. In either case, if the maximum is not achieved by an input distribution P satisfying $\Lambda_0(P) \geq \Lambda$, then the capacity (for deterministic codes) is strictly smaller than the random code capacity, while still being positive if the hypothesis of Theorem 3 part 2) (or of the Corollary part 2)) holds. This is illustrated by Example 2 in Section IV.

III. Proofs of Main Results

For notational convenience, joint types of length-n sequences will be represented by joint distributions of dummy random variables. Then if, for instance, X, S, Y represents a joint type, i.e., $P_{XSY} = P_{x,s,y}$ for some $x \in \mathcal{X}^n$, $s \in \mathcal{S}^n$, $y \in \mathcal{Y}^n$, we write $\tau_X = \{x\colon x \in \mathcal{X}^n,\ P_x = P_X\}$, $\tau_{XY} = \{(x, y)\colon x \in \mathcal{X}^n,\ y \in \mathcal{Y}^n,\ P_{x,y} = P_{XY}\}$, $\tau_{XSY} = \{(x, s, y)\colon x \in \mathcal{X}^n,\ s \in \mathcal{S}^n,\ y \in \mathcal{Y}^n,\ P_{x,s,y} = P_{XSY}\}$, etc. Similarly, we use self-explanatory notation for sections of τ_{XY}, τ_{XSY}, etc.; for example, $\tau_{Y|X}(x) = \{y\colon (x, y) \in \tau_{XY}\}$, $\tau_{Y|XS}(x, s) = \{y\colon (x, s, y) \in \tau_{XSY}\}$, etc.

We state below as facts a few basic bounds on types (cf., e.g., Csiszár–Körner [4]).

Fact 1: The number of possible joint types of sequences of length n is a polynomial in n.

Fact 2: We have

$$(n+1)^{-|\mathcal{X}|}\exp\{nH(X)\} \leq |\tau_X| \leq \exp\{nH(X)\}, \quad \text{if } \tau_X \neq \phi;$$

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|}\exp\{nH(Y|X)\} \leq |\tau_{Y|X}(x)| \leq \exp\{nH(Y|X)\}, \quad \text{if } \tau_{Y|X}(x) \neq \phi.$$

Fact 3: For any channel $V\colon \mathcal{X} \to \mathcal{Y}$,

$$\sum_{y\in\tau_{Y|X}(x)} V^n(y|x) \leq \exp\{-nD(P_{XY}\|P_X \times V)\}$$

where $P_X \times V$ denotes the distribution on $\mathcal{X} \times \mathcal{Y}$ with probability mass function $P_X(x)V(y|x)$.

The set of codewords $x_1, \cdots, x_N$ used in proving our main results is any set with the properties stated in the following lemma. We prove in the Appendix that a randomly selected codeword set will possess these properties with probability arbitrarily close to 1. We remark that Lemma 3 does not require the codewords to be distinct (although in our actual application this could be assumed).

Lemma 3: For any $\epsilon > 0$, $n \geq n_0(\epsilon)$, $N \geq \exp(n\epsilon)$, and type P, there exist codewords $x_1, \cdots, x_N$ in $\mathcal{X}^n$, each of type P, such that for every $x \in \mathcal{X}^n$, $s \in \mathcal{S}^n$, and every

joint type $P_{XX'S}$, upon setting $R = 1/n \log N$, we have

$$\left|\{j: (x, x_j, s) \in \tau_{XX'S}\}\right| \leq \exp\left\{n\left(\left|R - I(X' \wedge XS)\right|^+ + \epsilon\right)\right\} \quad (3.1)$$

$$\frac{1}{N}\left|\{i: (x_i, s) \in \tau_{XS}\}\right| \leq \exp(-n\epsilon/2), \quad \text{if } I(X \wedge S) > \epsilon \quad (3.2)$$

and

$$\frac{1}{N}\left|\{i: (x_i, x_j, s) \in \tau_{XX'S} \text{ for some } j \neq i\}\right| \leq \exp(-n\epsilon/2), \quad \text{if } I(X \wedge X'S) - \left|R - I(X' \wedge S)\right|^+ > \epsilon. \quad (3.3)$$

In addition to Lemma 3, we will need the following lemma which establishes the inambiguity of the decoding rule in Definition 3.

Lemma 4: If the AVC is nonsymmetrizable and $\beta > 0$, then for a sufficiently small η, no quintuple of random variables X, X', S, S', Y can simultaneously satisfy

$$P_X = P_{X'} = P \quad \text{with } \min_{x \in \mathcal{X}} P(x) \geq \beta \quad (3.4)$$

$$P_{XSY} \in \mathcal{C}_\eta, \qquad P_{X'S'Y} \in \mathcal{C}_\eta \quad (3.5)$$

and

$$I(XY \wedge X'|S) \leq \eta \qquad I(X'Y \wedge X|S') \leq \eta. \quad (3.6)$$

Proof: By the definition of $\mathcal{C}_\eta$ in (2.4), the condition $P_{XSY} \in \mathcal{C}_\eta$ means that

$$D(P_{XSY} \| P_X \times P_S \times W) = \sum_{x,s,y} P_{XSY}(x,s,y) \log \frac{P_{XSY}(x,s,y)}{P_X(x) P_S(s) W(y|x,s)} \leq \eta.$$

Upon adding to this

$$I(XY \wedge X'|S) = \sum_{x,s',s,y} P_{XX'SY}(x,x',s,y) \log \frac{P_{X'|XSY}(x'|x,s,y)}{P_{X'|S}(x'|s)} \leq \eta,$$

we obtain

$$\sum_{x,x',s,y} P_{XX'SY}(x,x',s,y) \cdot \log \frac{P_{XX'SY}(x,x',s,y)}{W(y|x,s) P(x) P_{X'S}(x',s)} \leq 2\eta.$$

Here, the left-hand side is the divergence of two distributions on $\mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, namely, of $P_{XX'SY}$ and the distribution with probability mass function $W(y|x,s)P(x)P(x')P_{S|X'}(s|x')$. Projecting both these distributions on $\mathcal{X} \times \mathcal{X} \times \mathcal{Y}$, the divergence does not increase. Hence we get

$$D(P_{XX'Y} \| P \times P \times V') \leq 2\eta \quad (3.7)$$

where $P \times P \times V'$ is defined by the probability mass function $P(x)P(x')V'(y|x,x')$, with

$$V'(y|x,x') = \sum_s W(y|x,s) P_{S|X'}(s|x'). \quad (3.8)$$

Since the variational distance between any two probability distributions is bounded above by the square root of their divergence times an absolute constant c (Pinsker's inequality [10], cf. Csiszár–Körner [4, p. 58]), it follows that

$$\sum_{x,x'y} \left|P_{XX'Y}(x,x',y) - P(x)P(x')V(y|x,x')\right| \leq c\sqrt{2\eta}. \quad (3.9)$$

Commencing with the conditions $P_{X'S'Y} \in \mathcal{C}_\eta$ and $I(X'Y \wedge X|S') \leq \eta$, we obtain in a similar manner that

$$\sum_{x,x',y} \left|P_{XX'Y}(x,x',y) - P(x)P(x')V(y|x,x')\right| \leq c\sqrt{2\eta} \quad (3.10)$$

where

$$V(y|x,x') = \sum_s W(y|x',s) P_{S|X}(s|x). \quad (3.11)$$

Comparing (3.9) and (3.10), we obtain that

$$\sum_{x,x',y} P(x)P(x')\left|V(y|x,x') - V'(y|x,x')\right| \leq 2c\sqrt{2\eta}$$

whence

$$\max_{x,x',y} \left|V(y|x,x') - V'(y|x,x')\right| \leq \frac{2c\sqrt{2\eta}}{\beta^2} \quad (3.12)$$

if $\min_{x \in \mathcal{X}} P(x) \geq \beta$.

For a nonsymmetrizable AVC, some $\xi > 0$ exists such that

$$\max_{x,x',y} \left|\sum_s W(y|x,s)U_1(s|x') - \sum_s W(y|x',s)U_2(s|x)\right| \geq \xi \quad (3.13)$$

for every $U_1: \mathcal{X} \to \mathcal{S}$, $U_2: \mathcal{X} \to \mathcal{S}$ (cf. Lemma A2 in the Appendix). In particular, for the choice of $U_1 = P_{S|X'}$, $U_2 = P_{S|X}$, (3.13) yields

$$\max_{x,x',y} \left|V(y|x,x') - V'(y|x,x')\right| \geq \xi, \quad (3.14)$$

which contradicts (3.12) if

$$\eta < \frac{\xi^2 \beta^4}{8c^2}.$$

Proof of Theorem 1: The necessity of nonsymmetrizability for $C > 0$ is well-known. In fact, as Ericson [9] indicates, an idea of Blackwell *et al.* [3] leads to the conclusion that for a symmetrizable AVC, every code with $N \geq 2$ codewords has

$$\max_{s \in \mathcal{S}^n} \bar{e}(s) \geq \frac{N-1}{2N}. \quad (3.15)$$

More specifically, this follows from (3.29) in the proof of Lemma 1 below. It is also well-known (and is a consequence of Lemma 2 with $\Lambda = l_{\max}$) that $C \leq \max_P I(P)$.

As observed after the statement of Theorem 1, nonsymmetrizability implies that $I(P) > 0$ for every strictly positive P. Thus it remains to establish the hard part of Theorem 1, namely, that for a nonsymmetrizable AVC, $\max_P I(P)$ is an achievable rate. To this end, since $I(P)$ is

a continuous function of P, it suffices to prove the following Lemma 5, which is our first key result.

Lemma 5: Given any nonsymmetrizable AVC and arbitrary $\beta > 0$, $\delta > 0$, for any block length $n \geq n_0$ and any type P with $\min P(x) > \beta$, there exists a code with codewords $x_1, \cdots, x_N$, each of type P, such that

$$\frac{1}{n} \log N > I(P) - \delta, \qquad \max_{s \in \mathcal{S}^n} \bar{e}(s) < \exp(-n\gamma). \quad (3.16)$$

Here n_0 and $\gamma > 0$ depend only on the given AVC, and on β and δ.

Proof: Let $x_1, \cdots, x_N$ be as in Lemma 3, with $R = 1/n \log N$ satisfying

$$I(P) - \delta < R < I(P) - \frac{2}{3}\delta \quad (3.17)$$

and with ϵ (from Lemma 3) to be specified later. Let the decoder ϕ be as described in Definition 3. Lemma 4 provides that this ϕ is unambiguously defined if η is chosen sufficiently small. In fact, if for some $y \in \mathcal{Y}^n$ and some $i \neq j$, both x_i and x_j satisfied conditions 1) and 2) in Definition 3, then some s and s' would exist, with the joint types of (x_i, x_j, s, s', y) being represented by the dummy random variables X, X', S, S', Y (i.e., $(x_i, x_j, s, s', y) \in \tau_{XX'SS'Y}$) that satisfy (3.4), (3.5), and (3.6) simultaneously. This contradicts Lemma 4.

To establish (3.16), fix any $s \in \mathcal{S}^n$, and observe first by (3.2) and Fact 1 that

$$\begin{aligned} &\frac{1}{N}\left|\left\{i: (x_i, s) \in \bigcup_{I(X \wedge S) > \epsilon} \tau_{XS}\right\}\right| \\ &\quad \leq (\text{number of joint types}) \cdot \exp(-n\epsilon/2) \\ &\quad \leq \exp(-n\epsilon/3). \end{aligned} \quad (3.18)$$

(All bounds in this proof are valid for n larger than a suitable threshold n_0, which depends on ϵ.)

Hence to obtain an exponentially decreasing upper bound on

$$\bar{e}(s) = \frac{1}{N}\sum_{i=1}^{N} e(i, s) = \frac{1}{N}\sum_{i=1}^{N} \sum_{y:\, \phi(y) \neq i} W^n(y|x_i, s), \quad (3.19)$$

it suffices to deal with only those codewords x_i for which $(x_i, s) \in \tau_{XS}$ with $I(X \wedge S) \leq \epsilon$. Then, for $P_{XSY} \notin \mathcal{G}_\eta$ (cf. (2.4)), we have

$$\begin{aligned} D(P_{XSY} \| P_{XS} \times W) &= D(P_{XSY} \| P_X \times P_S \times W) - I(X \wedge S) \\ &> \eta - \epsilon, \end{aligned}$$

and thus by Fact 3,

$$\begin{aligned} \sum_{y \in \tau_{Y|XS}(x_i, s)} W^n(y|x_i, s) &\leq \exp\{-nD(P_{XSY} \| P_{XS} \times W)\} \\ &< \exp\{-n(\eta - \epsilon)\}. \end{aligned}$$

Hence by Fact 1,

$$\sum_{y:\, P_{x_i, s, y} \in \mathcal{G}_\eta} W^n(y|x_i, s) \leq \exp\{-n(\eta - 2\epsilon)\}. \quad (3.20)$$

Next notice that if $P_{x_i, s, y} \in \mathcal{G}_\eta$ and $\phi(y) \neq i$, then condition 2) of Definition 3 must be violated. Let us therefore denote by $\mathcal{D}_\eta$ the set of all joint distributions $P_{XX'SY}$ such that 1) $P_{XSY} \in \mathcal{G}_\eta$; 2) $P_{X'S'Y} \in \mathcal{G}_\eta$ for some S'; and 3) $I(XY \wedge X'|S) > \eta$. Then it follows that

$$\sum_{\substack{y:\, (x_i, s, y) \in \mathcal{G}_\eta \\ \phi(y) \neq i}} W^n(y|x_i, s) \leq \sum_{P_{XX'SY} \in \mathcal{D}_\eta} e_{XX'SY}(i, s) \quad (3.21)$$

where

$$e_{XX'SY}(i, s) = \sum_{\substack{y:\, (x_i, x_j, s, y) \in \tau_{XX'SY} \\ \text{for some } j \neq i}} W^n(y|x_i, s) \quad (3.22)$$

and the summation in (3.21) extends to all joint types $P_{XX'SY} \in \mathcal{D}_\eta$ (of course, $e_{XX'SY}(i, s) = 0$ unless $P_{X'} = P_X = P$ and $P_{XS} = P_{x_i, s}$).

Combining (3.18)–(3.21), we have thus far obtained that

$$\begin{aligned} \bar{e}(s) \leq \exp(-n\epsilon/3) &+ \exp\{-n(\eta - 2\epsilon)\} \\ &+ \frac{1}{N}\sum_{i=1}^{N} \sum_{P_{XX'SY} \in \mathcal{D}_\eta} e_{XX'SY}(i, s). \end{aligned} \quad (3.23)$$

Before proceeding to bound $e_{XX'SY}(i, s)$, we notice that it suffices to do so when $P_{XX'SY} \in \mathcal{D}_\eta$ satisfies

$$I(X \wedge X'S) \leq |R - I(X' \wedge S)|^+ + \epsilon. \quad (3.24)$$

Otherwise, by (3.3),

$$\frac{1}{N}\left|\{i: (x_i, x_j, s) \in \tau_{XX'S} \text{ for some } j \neq i\}\right| < \exp(-n\epsilon/2).$$

Since $(x_i, x_j, s) \in \tau_{XX'S}$ for some $j \neq i$ is a necessary condition for $e_{XX'SY}(i, s) > 0$ (cf. (3.22)), it follows from Fact 1 that the contribution to the double summation in (3.23) of the terms with $P_{XX'SY} \in \mathcal{D}_\eta$ not satisfying (3.24) is less than $\exp(-\eta\epsilon/3)$.

Now from (3.22),

$$\begin{aligned} e_{XX'SY}(i, s) \leq \sum_{j:\, (x_i, x_j, s) \in \tau_{XX'S}} &\sum_{y \in \tau_{Y|XX'S}(x_i, x_j, s)} \\ &\cdot W^n(y|x_i, s). \end{aligned} \quad (3.25)$$

As $W^n(y|x_i, s)$ is constant for $y \in \tau_{Y|XS}(x_i, s)$ and this constant is less than or equal to $(|\tau_{Y|XS}(x_i, s)|)^{-1}$, the inner sum in (3.25) is bounded above by $|\tau_{Y|XX'S}(x_i, x_j, s)| \cdot (|\tau_{Y|XS}(x_i, s)|)^{-1}$, which in turn is less than or equal to $\exp\{-n(I(Y \wedge X'|XS) - \epsilon)\}$ by Fact 2. Hence using (3.1), it follows from (3.25) that

$$\begin{aligned} e_{XX'SY}(i, s) \leq \exp\Big\{-n\Big[&I(Y \wedge X'|XS) \\ &- |R - I(X' \wedge XS)|^+ - 2\epsilon\Big]\Big\}. \end{aligned} \quad (3.26)$$

To further bound $e_{XX'SY}(i, s)$ when (3.24) holds, we distinguish between two cases: a) $R \leq I(X' \wedge S)$, and b) $R > I(X' \wedge S)$.

In case a), (3.24) yields

$$I(X \wedge X'|S) \leq I(X \wedge X'S) \leq \epsilon,$$

and hence, by condition 3) in the definition of $\mathcal{D}$,

$$I(Y \wedge X'|XS) = I(XY \wedge X'|S) - I(X \wedge X'|S) \geq \eta - \epsilon.$$

Since now $R \le I(X' \wedge S) \le I(X' \wedge XS)$, it follows from (3.26) that

$$e_{XX'SY}(i,s) \le \exp(-n(\eta - 3\epsilon)). \tag{3.27}$$

In case b) we obtain from (3.24) that

$$\begin{aligned} R &> I(X \wedge X'S) + I(X' \wedge S) - \epsilon \\ &= I(X' \wedge XS) + I(X \wedge S) - \epsilon \\ &\ge I(X' \wedge XS) - \epsilon \end{aligned}$$

and hence

$$|R - I(X' \wedge XS)|^+ \ge R - I(X' \wedge XS) - \epsilon.$$

Substituting this into (3.26), it follows that for case b),

$$\begin{aligned} e_{XX'SY}(i,s) &\le \exp\{-n(I(X' \wedge XSY) - R - 3\epsilon)\} \\ &\le \exp\{-n(I(X' \wedge Y) - R - 3\epsilon)\}. \end{aligned} \tag{3.28}$$

Recall that $P_{XX'SY} \in \mathscr{D}_\eta$ implies, in particular, that $P_{X'S'Y} \in \mathscr{C}_\eta$ for some S'. Thus by the definition of $\mathscr{C}_\eta$ (cf. (2.4)), $P_{X'S'Y}$ is arbitrarily close to $P_{X''S''Y''} \in \mathscr{C}_0$ defined by $P_{X''S''Y''} = P \times P_{S'} \times W$, if η is sufficiently small; then $I(X' \wedge Y)$ is arbitrarily close to $I(X'' \wedge Y'')$, say, $I(X' \wedge Y) \ge I(X'' \wedge Y'') - \delta/3$. Using the definition (2.6) of $I(P)$ and the assumption in (3.17), it follows that

$$I(X' \wedge Y) - R \ge I(P) - \delta/3 - R \ge \delta/3$$

if η is sufficiently small and depends only on δ. Fixing the heretofore unspecified η accordingly (and small enough for the decoding rule to be unambiguous), (3.28) yields for case b) that

$$e_{XX'SY}(i,s) \le \exp\left\{-n\left(\frac{\delta}{3} - 3\epsilon\right)\right\}.$$

By the observation made in the paragraph containing (3.24), we obtain from (3.23) upon using (3.27), the previous bound, and Fact 1 that

$$\bar{e}(s) \le \exp(-n\epsilon/4)$$

if, for instance, $\epsilon \le \min(\eta/4, \delta/10)$ and n is sufficiently large. As the bound holds uniformly in $s \in \mathscr{S}^n$, the proof of Lemma 5, and thereby also of Theorem 1, is complete.

Proof of Lemma 1: Consider any code with codeword set $x_1, \cdots, x_N$ and decoder ϕ, where $x_i = (x_{i1}, \cdots, x_{in})$, $i = 1, \cdots, N$. For $U \in \mathscr{U}$, i.e., U satisfying (2.7), consider N $\mathscr{S}^n$-valued random variables $\boldsymbol{S}_j = (S_{j1}, \cdots, S_{jn})$ with statistically independent components, where $\Pr\{S_{jk} = s\} = U(s|x_{jk})$. Then for each pair (i,j) and every $y = (y_1, \cdots, y_n)$ in $\mathscr{Y}^n$, we have by (1.1) and the definition of $\boldsymbol{S}_j$ that

$$\begin{aligned} EW^n(y|x_i, \boldsymbol{S}_j) &= \prod_{k=1}^n EW(y_k|x_{ik}, S_{jk}) \\ &= \prod_{k=1}^n \sum_{s \in \mathscr{S}} W(y_k|x_{ik}, s) U(s|x_{jk}). \end{aligned}$$

On account of (2.7), it follows that

$$EW^n(y|x_i, \boldsymbol{S}_j) = EW^n(y|x_j, \boldsymbol{S}_i),$$

and hence by (2.1), for $i \ne j$, we have

$$\begin{aligned} &Ee(i, \boldsymbol{S}_j) + Ee(j, \boldsymbol{S}_i) \\ &\quad = \sum_{y:\, \phi(y) \ne i} EW^n(y|x_i, \boldsymbol{S}_j) + \sum_{y:\, \phi(y) \ne j} EW^n(y|x_j, \boldsymbol{S}_i) \\ &\quad \ge \sum_{y \in \mathscr{Y}^n} EW^n(y|x_i, S_j) = 1. \end{aligned}$$

Using this fact and (2.2), we obtain

$$\begin{aligned} \frac{1}{N}\sum_{j=1}^N E\bar{e}(\boldsymbol{S}_j) &= \frac{1}{N^2}\sum_{i=1}^N \sum_{j=1}^N E\bar{e}(i, \boldsymbol{S}_j) \\ &\ge \frac{1}{N^2} \cdot \frac{N(N-1)}{2} = \frac{N-1}{2N}, \end{aligned}$$

whence it follows that

$$E\bar{e}(\boldsymbol{S}_j) \ge \frac{N-1}{2N} \ge \frac{1}{4}, \qquad \text{for some } j \in \{1, \cdots, N\}. \tag{3.29}$$

Suppose now that each codeword x_j is of type P where $\Lambda_0(P) < \Lambda$, and let $U \in \mathscr{U}$ attain the minimum in (2.13). Then using (2.10),

$$\begin{aligned} El(\boldsymbol{S}_j) &= \frac{1}{n}\sum_{k=1}^n El(S_{jk}) = \frac{1}{n}\sum_{k=1}^n \sum_{s \in \mathscr{S}} l(s) U(s|x_{jk}) \\ &= \sum_{x \in \mathscr{X}} \sum_{s \in \mathscr{S}} P(x) U(s|x) l(s) = \Lambda_0(P) \end{aligned}$$

and

$$\operatorname{var} l(\boldsymbol{S}_j) = \frac{1}{n^2}\sum_{k=1}^n \operatorname{var} l(S_{jk}) \le \frac{l_{\max}^2}{n}.$$

Hence by Chebyshev's inequality,

$$\begin{aligned} \Pr\{l(\boldsymbol{S}_j) > \Lambda\} &= \Pr\{l(\boldsymbol{S}_j) - El(\boldsymbol{S}_j) > \Lambda - \Lambda_0(P)\} \\ &\ge \frac{l_{\max}^2}{n}(\Lambda - \Lambda_0(P))^{-2}. \end{aligned} \tag{3.30}$$

Since

$$E\bar{e}(\boldsymbol{S}_j) \le \max_{s:\, l(s) \le \Lambda} \bar{e}(s) + \Pr\{l(\boldsymbol{S}_j) > \Lambda\},$$

the lemma follows from (3.29) and (3.30).

Proof of Lemma 2: First we show that for some S satisfying

$$El(S) \le \Lambda(1 - \eta) \tag{3.31}$$

(with $\eta > 0$ depending on δ but not on P) and for

$$P_{XSY}(x,s,y) = P(x) P_S(s) W(y|x,s), \tag{3.32}$$

we have

$$I(X \wedge Y) \le I(P, \Lambda) + \delta/2. \tag{3.33}$$

In fact, let P_{S^*} achieve the minimum in (2.14), i.e., let $I(X \wedge Y^*) = I(P, \Lambda)$ for $P_{XS^*Y^*}$ as in (3.32) with $El(S^*) \le \Lambda$. Pick $s_0 \in \mathscr{S}$ with $l(s_0) = 0$ (cf. (2.11)), and define $P_S(s) = (1 - \eta) P_{S^*}(s)$ if $s \ne s_0$, $P_S(s_0) = \eta + (1 - \eta) P_{S^*}(s_0)$. Then (3.31) is clearly satisfied, and so is (3.33)

for a sufficiently small η, since $I(X \wedge Y)$ is a uniformly continuous function of the pair (P, P_S) if P_{XSY} is given by (3.32).

Now consider any code with codewords $x_1, \cdots, x_N$ and decoder ϕ, and let $\boldsymbol{S} = (S_1, \cdots, S_n)$ be n independent repetitions of S as defined earlier. Then by (2.1), (2.2), (1.1), and the independence of $S_1, \cdots, S_n$,

$$E\bar{e}(\boldsymbol{S}) = \frac{1}{N}\sum_{i=1}^{N} E(i, \boldsymbol{S}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{y:\, \phi(y)\neq i} EW^n(y|x_i, \boldsymbol{S})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{y:\, \phi(y)\neq i}\prod_{j=1}^{n} EW(y_j|x_{ij}, S_j). \tag{3.34}$$

Introducing a discrete memoryless channel (DMC) $\{W_S\}$ defined by

$$W_S(y|x) = EW(y|x, S), \tag{3.35}$$

(3.34) yields that $E\bar{e}(\boldsymbol{S})$ is equal to $\bar{e}_{W_S}$, the average probability of error when the given code is used on the DMC $\{W_S\}$. Since (3.31) implies, by (2.10) and Chebyshev's inequality, that

$$\Pr\{l(\boldsymbol{S}) > \Lambda\} = \Pr\left\{\frac{1}{n}\sum_{i=1}^{n} l(S_i) > El(S) + \eta\Lambda\right\}$$

$$\leq \frac{\{\operatorname{var} l(S)\}^2}{n(\eta\Lambda)^2} \leq \frac{l_{\max}^2}{n\eta^2\Lambda^2},$$

it follows that

$$\max_{s:\, l(s)\leq\Lambda} \bar{e}(s) \geq E\bar{e}(\boldsymbol{S}) - \Pr\{l(\boldsymbol{S}) > \Lambda\}$$

$$\geq \bar{e}_{W_S} - \frac{l_{\max}^2}{n\eta^2\Lambda^2}. \tag{3.36}$$

Finally, notice that (3.32) means that Y is connected with X by the channel W_S defined in (3.35). Hence (3.33) implies, by the strong converse to the coding theorem for a DMC with codewords of type P (cf. Csiszár–Körner [4, corollary 1.4, p. 104]), that if all the codewords $x_1, \cdots, x_N$ are of type P, then $\bar{e}_{W_S}$ is arbitrarily close to 1 if $1/n \log N \geq I(P, \Lambda) + \delta$ and n is large enough. This and (3.36) complete the proof of Lemma 2.

Proof of Theorem 2: As mentioned in Section II, we now use a slightly modified decoding rule, replacing $\mathscr{C}_\eta$ in Definition 3 by

$$\mathscr{C}_\eta(\Lambda) = \{P_{XSY}\colon P_{XSY} \in \mathscr{C}_\eta,\ El(S) \leq \Lambda\}.$$

To prove that this modified decoding rule is unambiguous if the codewords are of type P satisfying (2.15), we have to establish that no quintuple of random variables X, X', S, S', Y can simultaneously satisfy

$$P_X = P_{X'} = P \qquad \text{with } \Lambda_0(P) \geq \Lambda + \alpha,\ \min_{x\in\mathscr{X}} P(x) \geq \beta \tag{3.4'}$$

$$P_{XSY} \in \mathscr{C}_\eta(\Lambda), \qquad P_{X'S'Y} \in \mathscr{C}_\eta(\Lambda), \tag{3.5'}$$

and (3.6). The proof is identical to that of Lemma 4, with the only difference being that now (3.13) need not hold for every U_1, U_2. This does not, however, affect the proof because by the second assertion of Lemma A2, (3.13) does hold subject to the constraint (A.14) which, by assumptions (3.4') and (3.5'), is satisfied for $U_1 = P_{S|X'}$, $U_2 = P_{S|X}$.

Using the codeword set of Lemma 3 and the decoder specified earlier, the remainder of the proof of Theorem 2 is identical to that of Lemma 5 and is, therefore, omitted.

Proof of Theorem 3: Part 1) follows immediately from Lemma 1. To prove part 2), we first claim that

$$F(\alpha) = \max_{\substack{P:\, g(P)\leq\Gamma-\alpha \\ \Lambda_0(P)\geq\Lambda+\alpha}} I(P, \Lambda) \tag{3.37}$$

is a continuous function of α in a sufficiently small neighborhood of 0, say $(-\eta, \eta)$. To see this, observe that by (2.14), $I(P, \Lambda)$ is the minimum of a family of concave functions of P (since $I(X \wedge Y)$ is concave in P for $P_{XSY} = P \times P_S \times W$ for a fixed P_S) and hence is itself a concave function of P. Similarly, $\Lambda_0(P)$ is also a concave function of P. It then follows in a standard manner that $F(\alpha)$ is a concave function of α in the interval where $\{P\colon g(P) \leq \Gamma - \alpha,\ \Lambda_0(P) \geq \Lambda + \alpha\} \neq \phi$. The assumption $\max_{P:\, g(P)\leq\Gamma} \Lambda_0(P) > \Lambda$ implies that this interval contains 0 in its interior, establishing our first claim.

Now for any $\alpha > 0$, let P^* achieve the maximum in (3.37), and let $\xi > 0$ be small enough so that for every P with

$$\max_{x\in\mathscr{X}} |P(x) - P^*(x)| < \xi, \tag{3.38}$$

$I(P, \Lambda)$ differs from $I(P^*, \Lambda) = F(\alpha)$ by less than δ, and $g(P) \leq \Gamma$, $\Lambda_0(P) \geq \Lambda + \alpha/2$. If $\beta > 0$ is small enough, there exists (for sufficiently large n) a type P satisfying (3.38), as also $\min_{x\in\mathscr{X}} P(x) \geq \beta$. Then by Theorem 2 a code exists with codewords $x_1, \cdots, x_N$, each of type P—thus satisfying $g(x_i) \leq \Gamma$—such that

$$\frac{1}{n}\log N > I(P, \Lambda) - \delta > F(\alpha) - 2\delta$$

and $\max_{s:\, l(s)\leq\Lambda} \bar{e}(s)$ is as small as desired. Since $F(\alpha)$ is continuous at $\alpha = 0$, this proves the forward part of Theorem 3, i.e., the achievability of $R = F(0)$ under input constraint Γ and state constraint Λ (cf. Definition 4).

To prove the converse, i.e., that no $R > F(0)$ is an achievable rate, observe that Lemmas 1 and 2 immediately imply that no R larger than

$$\max_{\substack{P:\, g(P)\leq\Gamma \\ \Lambda_0(P)\geq\Lambda-\alpha}} I(P, \Lambda)$$

can be achieved for any $\alpha > 0$. Since the last maximum $\leq F(-\alpha)$ (cf. (3.27)), the desired converse follows from the continuity of F at $\alpha = 0$.

Remark: While the case $\max_{P:\, g(P)\leq\Gamma} \Lambda_0(P) = \Lambda$ remains unsolved in general, for certain AVC's it is easily seen that $C(\Gamma, \Lambda) = 0$ in this case, too. This occurs if the set $\mathscr{U}$ of all channels satisfying (2.7) is the convex hull of a set of deterministic channels, i.e., of 0–1 matrices. In fact, the minimum in (2.13) is then attained for some 0–1 matrix U, and in the proof of Lemma 1 the state sequences

$\boldsymbol{S}_j$ will now be nonrandom, say $\boldsymbol{S}_j = s_j$. Then (3.29) reduces to $\bar{e}(s_j) \geq (N-1)/2N$, where $l(s_j) = \Lambda_0(P)$ (rather than $El(\boldsymbol{S}_j) = \Lambda_0(P)$). Thus codes with codewords of type P cannot be "good" under state constraint Λ, even when $\Lambda_0(P) \leq \Lambda$.

IV. Examples

Two simple examples are considered in this section. In both examples, the input alphabet and the set of states are binary, i.e., $\mathcal{X} = \mathcal{S} = \{0,1\}$, and the channel output is a deterministic function of the input and the state. Furthermore, the functions $g(x) = x$ and $l(s) = s$ are used in the input and state constraints in either example. Thus $g(\boldsymbol{x})$ and $l(\boldsymbol{s})$ are the respective normalized Hamming weights of the binary n-sequences $\boldsymbol{x}$ and $\boldsymbol{s}$.

Example 1: Let $\mathcal{Y} = \{0,1\}$ and let $W(y|x,s) = 1$ if $y = x + s$ modulo 2, and 0 otherwise. This is a symmetric AVC, and hence $C = 0$. Further, since $P_{XSY} \in \mathcal{C}_0$ (cf. (2.5)) if and only if

$$Y = X + S \bmod 2, \qquad X \text{ and } S \text{ independent}, \quad (4.1)$$

we obtain that $I(P) = 0$ for every P (cf. (2.6)), as $P_S = (1/2, 1/2)$ in (4.1) yields $I(X \wedge Y) = 0$. Thus the random code capacity of this AVC is also equal to zero.

To determine the capacity $C(\Gamma, \Lambda)$ under input constraint Γ and state constraint Λ, we first evaluate $\Lambda_0(P)$ (cf. (2.13)). For this AVC, (2.7) is satisfied if and only if U is symmetric, i.e.,

$$\mathcal{U} = \left\{ \begin{pmatrix} 1-u & u \\ u & 1-u \end{pmatrix} : 0 \leq u \leq 1 \right\},$$

and thus for $P = (1-p, p)$, (2.13) yields

$$\Lambda_0(P) = \min_{0 \leq u \leq 1} [(1-p)u + p(1-u)] = \min(p, 1-p). \quad (4.2)$$

By Theorem 3 part 1), and the remark following the proof of Theorem 3, we have $C(\Gamma, \Lambda) = 0$ if $\max_{P:\, g(P) \leq \Gamma} \Lambda_0(P) \leq \Lambda$. Since $g(P) = p$, by (4.2) we get

$$C(\Gamma, \Lambda) = 0, \qquad \text{if } \Lambda \geq \min(\Gamma, 1/2). \quad (4.3)$$

For $\Lambda < \min(\Gamma, 1/2)$, $C(\Gamma, \Lambda)$ is given by Theorem 3 part 2). To determine it explicitly, we must find $I(P, \Lambda)$ (cf. (2.14)).

For $P_X = P = (1-p, p)$, $P_S = (1-q, q)$ in (4.1), write

$$I_1(p,q) = I(X \wedge Y) = H(Y) - H(Y|X)$$
$$= h(p * q) - h(q) \quad (4.4)$$

where $p * q = pq + (1-p)(1-q)$, and $h(t) = -t \log t - (1-t)\log(1-t)$ is the binary entropy function. By standard properties of mutual information (cf., e.g., Csiszár–Körner [4, p. 50, lemma 3.5 (d)]), $I_1(p,q)$ is concave in p and convex in q. For a fixed p, $I_1(p,q)$ is minimized when $q = 1/2$; hence it is a decreasing function of q for $0 \leq q \leq 1/2$. Since $El(S) = q$, it follows from (2.14) that

$$I(P, \Lambda) = \min_{q \leq \Lambda} I_1(p,q)$$
$$= \begin{cases} I_1(p, \Lambda), & \text{if } \Lambda < 1/2 \\ 0, & \text{if } \Lambda \geq 1/2. \end{cases} \quad (4.5)$$

Now Theorem 3 part 2) gives, by (4.2) and the fact $g(P) = p$,

$$C(\Gamma, \Lambda) = \max_{\Lambda \leq p \leq \Gamma} I_1(p, \Lambda), \qquad \text{if } \Lambda < \min(\Gamma, 1/2).$$

Since $I_1(p,q)$, defined in (4.4), is concave in p and maximized when $p = 1/2$ (for any fixed q), we finally obtain that

$$C(\Gamma, \Lambda) = \begin{cases} I_1(1/2, \Lambda) = 1 - h(\Lambda), & \text{if } \Lambda < 1/2 \leq \Gamma \\ I_1(\Gamma, \Lambda) = h(\Gamma * \Lambda) - h(\Lambda), & \\ & \text{if } \Lambda < \Gamma \leq 1/2. \end{cases} \quad (4.6)$$

Notice that the random coding capacity under input constraint Γ and state constraint Λ is, by (2.19) and (4.5),

$$C_r(\Gamma, \Lambda) = \begin{cases} 0, & \text{if } \Lambda \geq 1/2 \\ I_1(1/2, \Lambda), & \text{if } \Lambda < 1/2 \leq \Gamma \\ I_1(\Gamma, \Lambda), & \text{if } \Lambda < 1/2,\ \Gamma < 1/2. \end{cases} \quad (4.7)$$

Thus the capacity under input constraint Γ and state constraint Λ is equal to the corresponding random code capacity, except for the case $\Gamma \leq \Lambda < 1/2$ when $C(\Gamma, \Lambda) = 0$ while $C_r(\Gamma, \Lambda) > 0$.

As a particular case of (4.6), the capacity of the AVC in this example under state constraint $\Lambda < 1/2$ and with no input constraint equals $C(1, \Lambda) = 1 - h(\Lambda)$. A remarkable feature of this result is that this capacity is the same as that of a binary symmetric channel (BSC) with crossover probability Λ. Since the AVC in this example is deterministic, the error probability $e(i, \boldsymbol{s})$ (cf. (2.1)) is either 0 or 1 for any code, any message, and any state sequence. Hence $\bar{e}(\boldsymbol{s})$, defined in (2.2), is simply the average number of incorrectly decoded messages when the state (or "noise") sequence is $\boldsymbol{s}$. Notice that the expected value of this $\bar{e}(\boldsymbol{s})$ over the ensemble of "noise vectors" $\boldsymbol{s}$ with independent bits, each of which has probability Λ of being equal to 1, is just the average probability of error of the given code over a BSC with crossover probability Λ. By the standard coding theorem for a BSC, this can be made arbitrarily small while maintaining a rate close to $1 - h(\Lambda)$. The result we have established says that the same rate is achievable even under the stronger requirement that the fraction of incorrectly decoded messages be small not only in expected value over an ensemble of noise vectors $\boldsymbol{s}$ but for every $\boldsymbol{s}$ individually, subject to $l(\boldsymbol{s}) \leq \Lambda$ (where, of course, $\boldsymbol{s}$ is unknown both to the sender and decoder).

It is instructive to point out that, for the AVC in this example, the problem of determining the m-capacity (rather

than the a-capacity) under state constraint Λ is equivalent to a basic unsolved problem of coding theory. The m-capacity of an AVC under state constraint Λ is the largest R such that for every $\epsilon > 0$ and $\delta > 0$, codes exist with $1/n \log N > R - \delta$ and $\max_{s:\, l(s) \le \Lambda} \max_{1 \le i \le N} e(i,s) < \epsilon$. For a deterministic AVC the last condition means that $e(i,s) = 0$ for every message i (rather than for "most messages" as before) and for every s with $l(s) \le \Lambda$. Thus the m-capacity of the AVC in this example under state constraint Λ equals the maximum rate of binary codes such that the normalized Hamming distance of any two codewords is larger than 2Λ.

Example 2: Let $\mathscr{Y} = \{0,1,2\}$, and let $W(y|x,s) = 1$ if $y = x + s$, and 0, otherwise. This is the simplest example (due to Blackwell *et al.* [3]) of an AVC with $C = 0$ and positive random code capacity.

To determine $C(\Gamma, \Lambda)$, we may assume that $\Lambda < 1$. Since for this AVC, only U equal to the identity matrix satisfies (2.7), we obtain from (2.13) that $\Lambda_0(P) = p$ for any $P = (1-p, p)$. Also, since $g(P) = p$, by Theorem 3 and the remark following its proof we have

$$C(\Gamma, \Lambda) = \begin{cases} 0, & \text{if } \Gamma \le \Lambda \\ \max_{\Lambda \le p \le \Gamma} I(P, \Lambda), & \text{if } \Gamma > \Lambda. \end{cases} \tag{4.8}$$

Now observe that $P_{XSY} \in \mathscr{C}_0$ (cf. (2.5)) iff

$$Y = X + S, \qquad X \text{ and } S \text{ independent.} \tag{4.9}$$

Then

$$I(X \wedge Y) = I_2(p,q) = H(pq, (1-p)(1-q), p+q-2pq) - h(q) \tag{4.10}$$

if $P_X = P = (1-p, p)$, $P_S = (1-q, q)$, and from (2.14),

$$I(P, \Lambda) = \min_{q \le \Lambda} I_2(p,q). \tag{4.11}$$

As in Example 1, it is seen that $I_2(p,q)$ is concave in p and convex in q. Further, we can see by differentiation that $p^* = 1/2$, $q^* = 1/2$ is a saddle point of $I_2(p,q)$. Thus the random code capacity, without constraints, is

$$C_r = \max_p \min_q I_2(p,q) = I_2(p^*, q^*) = \frac{1}{2}.$$

The random code capacity with input constraint Γ and state constraint Λ is

$$C_r(\Gamma, \Lambda) = \max_{p \le \Gamma} I(P, \Lambda) = \max_{p \le \Gamma} \min_{q \le \Lambda} I_2(p,q), \tag{4.12}$$

and thus $C_r(\Gamma, \Lambda) = C_r = 1/2$ if $\Gamma \ge 1/2$, $\Lambda \ge 1/2$.

Observe that $I(P, \Lambda)$ is a concave function of p since by (4.11) it is the minimum of concave functions. If $\Lambda \ge 1/2$, this function is maximized at $p^* = 1/2$. Hence the maximum in (4.8) is attained at $p = \Lambda$ if $\Gamma > \Lambda > 1/2$. Then writing $P_\Lambda = (1-\Lambda, \Lambda)$, we get

$$C(\Gamma, \Lambda) = I(P_\Lambda, \Lambda) = \min_{q \le \Lambda} I_2(\Lambda, q) < C_r.$$

This means that $C(\Gamma, \Lambda)$ is positive but smaller than the corresponding random code capacity $C_r(\Gamma, \Lambda) = C_r = 1/2$ if $\Gamma > \Lambda > 1/2$. In particular, the capacity under state constraint $\Lambda > 1/2$ and no input constraint is positive but less than the corresponding random code capacity.

On the other hand, if $\Lambda \le 1/2$, $\Gamma > \Lambda$, then $C(\Gamma, \Lambda) = C_r(\Gamma, \Lambda)$. To verify this, we need only establish that the input distribution $\tilde{P} = (1-\tilde{p}, \tilde{p})$ achieving the maximum in (4.12) satisfies $\tilde{p} \ge \Lambda$. This will follow if we show that $I(P, \Lambda)$ is an increasing function of p in the interval $0 \le p \le 1/2$ if $\Lambda \le 1/2$; thus by (4.11) it suffices to show that $I_2(p,q)$ is an increasing function of $0 \le p \le 1/2$ if $q \le 1/2$. This follows by differentiation. In fact, as $I_2(p,q)$ is concave in p, it suffices to check $(\partial/\partial p) I_2(p,q)$ at $p = 1/2$, which is seen to be nonnegative if $q \le 1/2$.

V. Discussion

Ahlswede [2] demonstrated that the capacity C of an AVC for (deterministic codes and) average probability of error is equal either to its random code capacity or else to zero. A necessary and sufficient computable characterization of AVC's for deciding between these alternatives was not available. We have established that nonsymmetrizability, stated by Ericson [9] as a necessary condition for $C > 0$, is in fact both necessary and sufficient; for a nonsymmetrizable AVC, C equals its random code capacity. Our proof does not rely on Ahlswede's [2] theorem. A good codeword set is selected at random, using a bounding technique of Dobrushin–Stambler [8]. A subtle decoding rule, similar to that in Csiszár–Körner [5], leads to an adequate bound on error probability.

Employing the same method we have also determined the AVC capacity when constraints are imposed on the state sequences. Now symmetrizability need no longer render $C = 0$. Instead, the crucial factor is whether or not $\Lambda_0 = \max_P \Lambda_0(P)$ (cf. (2.13)) is larger than the state constraint Λ. As remarked in [7], Ahlswede's elimination technique [2] does not apply when state constraints are present unless the capacity without state constraints is positive. Our results demonstrate that under a state constraint the capacity may be positive but less than the corresponding random code capacity. Similar results have also been obtained for the case where constraints are additionally imposed on the transmitted codewords.

Our results resolve as a special case a weakened version of a fundamental problem of coding theory. This unsolved problem concerns the determination of the largest asymptotic rate of binary codes which enables a codeword to be correctly recovered, regardless of which error vector of normalized Hamming weight Λ is added to it (mod 2). If, instead, an arbitrarily small fraction of the codewords is allowed to be incorrectly decoded, we have shown that the largest achievable rate is equal to the capacity of a BSC with crossover probability Λ.

In this paper the input and state constraints are "peak constraints" in the sense of [7]. In [7], for random codes,

"average constraints" were also considered where "average state constraint Λ" meant that the state sequence could be random subject to $E((1/n)\sum_{i=1}^{n} l(S_i)) \le \Lambda$. The same could be done for deterministic codes, too. Indeed, for nonsymmetrizable AVC's the results are completely analogous to those in [7]. For symmetrizable AVC's, however, a difficulty in determining the ϵ-capacity under "average state constraint Λ" arises due to the fact that while a symmetrizable AVC has zero capacity, its ϵ-capacity may be nonzero for $\epsilon > 1/2$.

We conclude with a comment on another aspect of the relation of our work to that of Ahlswede [2]. Ahlswede has established by his elimination technique that the capacity of an AVC for codes with a stochastic encoder (and deterministic decoder) and maximum probability of error is the same as its capacity for deterministic codes and average probability of error criterion. We remark that even though the elimination technique does not apply in the presence of state constraints, the said result nevertheless remains true. To see this, a minor modification of our proof yields the existence of codeword sets, as in Theorem 2, with the additional property that the message set $\{1,\cdots,N\}$ can be partitioned into subsets $A_1,\cdots,A_{N'}$ of sizes $\simeq \exp(n\epsilon)$ such that for each subset,

$$\max_{s:\, l(s)\le\Lambda} \frac{1}{|A_k|} \sum_{i\in A_k} e(i,s) < \exp(-n\gamma).$$

The resulting code can then be modified to a code with a stochastic encoder and with a message set $\{1,\cdots,N'\}$. Each $k \in \{1,\cdots,N'\}$ is encoded by a codeword randomly selected from A_k. Clearly, this new code will have maximum probability of error less than $\exp(-n\gamma)$ for every $s \in \mathcal{S}^n$ with $l(s) \le \Lambda$.

APPENDIX

We now prove Lemma 3 and another technical lemma referred to in Section III. We will show that $N = \exp(nR)$ randomly selected codewords will possess, with probability close to 1, all the properties stated in Lemma 3. Inequalities (3.1) and (3.2) are a consequence of Csiszár–Körner [5, lemma 1]; nevertheless, for completeness we give a simple proof. To establish (3.3), Chernoff bounding has to be applied to dependent random variables as in Dobrushin–Stambler [8]. The Chernoff bound required by us is stated as Lemma A1, and is related to [8, lemma 9].

Lemma A1: Let $Z_1,\cdots,Z_N$ be arbitrary random variables, and let $f_i(Z_1,\cdots,Z_i)$ be arbitrary with $0 \le f_i \le 1$, $i=1,\cdots,N$. Then the condition

$$E[f_i(Z_1,\cdots,Z_i)|Z_1,\cdots,Z_{i-1}] \le a \text{ a.s.}, \qquad i=1,\cdots,N, \tag{A1}$$

implies that

$$\Pr\left\{\frac{1}{N}\sum_{i=1}^{N} f_i(Z_1,\cdots,Z_i) > t\right\} \le \exp\{-N(t-a\log e)\}. \tag{A2}$$

Of course, (A2) is a nontrivial bound only for $t > a \log e$.

Proof: We observe that

$$\begin{aligned}
&\Pr\left\{\frac{1}{N}\sum_{i=1}^{N} f_i(Z_1,\cdots,Z_i) > t\right\} \\
&\quad = \Pr\left\{\exp\sum_{i=1}^{N} f_i(Z_1,\cdots,Z_i) > \exp(Nt)\right\} \\
&\quad \le \exp(-Nt)\, E\left[\exp\sum_{i=1}^{N} f_i(Z_1,\cdots,Z_i)\right] \\
&\qquad\qquad \text{(by Markov's inequality)} \\
&\quad = \exp(-Nt)\, E\left[\left(\exp\sum_{i=1}^{N-1} f_i(Z_1,\cdots,Z_i)\right)\right. \\
&\qquad \left.\cdot E(\exp f_N(Z_1,\cdots,Z_N)|Z_1,\cdots,Z_{N-1})\right].
\end{aligned} \tag{A3}$$

Since $0 \le f \le 1$ implies that $\exp f \le 1+f$ (recall that exponentials are to the base 2), we obtain by assumption (A1) and (A3) that

$$\begin{aligned}
&E(\exp f_N(Z_1,\cdots,Z_N)|Z_1,\cdots,Z_{N-1}) \\
&\quad \le 1 + E(f_N(Z_1,\cdots,Z_N)|Z_1,\cdots,Z_{N-1}) \\
&\quad \le 1 + a \le e^a = \exp(a\log e).
\end{aligned} \tag{A4}$$

By substituting this into (A3) and repeating the procedure in (A3), (A4) $(N-1)$ times, we obtain (A2).

Proof of Lemma 3: Let $Z_1,\cdots,Z_N$ be independent random variables, each uniformly distributed on τ_X. First fix $x \in \tau_X$, $s \in \mathcal{S}^n$, and a joint type $P_{XX'S}$ with $P_{XS} = P_{x,s}$, $P_{X'} = P_{X'}$. Apply Lemma A1 to

$$f_j(Z_1,\cdots,Z_j) = \begin{cases} 1, & \text{if } Z_j \in \tau_{X'|XS}(x,s) \\ 0, & \text{otherwise} \end{cases} \tag{A5}$$

(as the random variables defined in (A5) are independent and identically distributed, the full strength of Lemma A1 is not needed at this point). By Fact 2 in Section III, the condition (A1) is now fulfilled with

$$\begin{aligned}
a = \Pr\{Z_j \in \tau_{X'|XS}(x,s)\} &= \frac{|\tau_{X'|XS}(x,s)|}{|\tau_X|} \\
&\le \frac{\exp\{nH(X'|XS)\}}{(n+1)^{-|\mathcal{X}|}\exp\{nH(X)\}} \\
&= (n+1)^{|\mathcal{X}|}\exp\{-nI(X'\wedge XS)\}
\end{aligned}$$

where the last step follows because $H(X') = H(X)$. Setting

$$t = \frac{1}{N}\exp\left\{n\left(|R - I(X'\wedge XS)|^+ + \epsilon\right)\right\}$$

where $R = n^{-1}\log N$, we see that $N(t - a\log e) \ge (1/2)\exp(n\epsilon)$ if $n \ge n_1(\epsilon)$, where

$$n_1(\epsilon) = \min\left\{n: (n+1)^{|\mathcal{X}|}\log e < \frac{1}{2}\exp(n\epsilon)\right\}. \tag{A6}$$

Then (A2) results in

$$\begin{aligned}
\Pr\left\{\left|\{j: Z_j \in \tau_{X'|XS}(x,s)\}\right| > \exp\left[n\left(|R - I(X'\wedge XS)|^+ + \epsilon\right]\right\}\right. \\
\left. < \exp\left[-\frac{1}{2}\exp(n\epsilon)\right]\right.
\end{aligned} \tag{A7}$$

By the same argument, replacing $\tau_{X'|XS}(x,s)$ by $\tau_{X'|S}(s)$ in (A5), we also get

$$\Pr\left\{\left|\{j: Z_j \in \tau_{X'|S}(s)\}\right| > \exp\left[n\left(|R - I(X' \wedge S)|^+ + \epsilon\right]\right\} < \exp\left[-\frac{1}{2}\exp(n\epsilon)\right]. \quad \text{(A8)}$$

In particular, if $I(X' \wedge S) \le \epsilon$ (and $R \ge \epsilon$ as postulated), from (A8) with $\epsilon/2$ replacing ϵ, we obtain for $n \ge n_1(\epsilon/2)$ that

$$\Pr\left\{\frac{1}{N}\left|\{j: Z_j \in \tau_{X'|S}(s)\}\right| > \exp\left(-\frac{n\epsilon}{2}\right)\right\} < \exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{2}\right)\right]. \quad \text{(A9)}$$

The doubly exponential bounds (A7) and (A9) will suffice to establish (3.1) and (3.2). To obtain (3.3), we proceed as follows. Let A_i denote the set of indices $j < i$ such that $z_j \in \tau_{X'|S}(s)$, provided their number does not exceed $\exp\{n(|R - I(X' \wedge S)|^+) + (\epsilon/4)\}$; else, let $A_i = \phi$. Further, let

$$f_i(z_1, \cdots, z_i) = \begin{cases} 1, & \text{if } z_i \in \bigcup_{j \in A_i} \tau_{X|X'S}(z_j, s); \\ 0, & \text{otherwise.} \end{cases} \quad \text{(A10)}$$

Then by (A8), applied with $\epsilon/4$ instead of ϵ and for $n \ge n_1(\epsilon/4)$, we have

$$\Pr\left\{\sum_{i=1}^{N} f_i(Z_1, \cdots, Z_i) \ne \left|\{j: Z_i \in \tau_{X|X'S}(Z_j, s) \text{ for some } j < i\}\right|\right\} < \exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{4}\right)\right]. \quad \text{(A11)}$$

By the independence of the Z_i and by Fact 2, we obtain from (A10) that

$$\begin{aligned} &E\left(f_i(Z_1, \cdots, Z_i) | Z_1, \cdots, Z_{i-1}\right) \\ &= \Pr\left\{Z_i \in \bigcup_{j \in A_i} \tau_{X|X'S}(Z_j, s) \middle| Z_1, \cdots, Z_{i-1}\right\} \\ &\le |A_i| \frac{\exp\{nH(X|X'S)\}}{(n+1)^{-|\mathcal{X}|}\exp\{nH(X)\}} \\ &\le (n+1)^{|\mathcal{X}|}\exp\left\{n\left(|R - I(X' \wedge S)|^+ - I(X \wedge X'S) + \frac{\epsilon}{4}\right)\right\}. \end{aligned}$$

Supposing that $I(X \wedge X'S) > |R - I(X' \wedge S)|^+ + \epsilon$, (A1) holds with

$$a = (n+1)^{|\mathcal{X}|}\exp\left(-\frac{3}{4}n\epsilon\right).$$

Then (A2), with $t = \exp(-n\epsilon/2)$ and for $n \ge n_1(\epsilon/4)$ (cf. (A6)), yields that

$$\Pr\left\{\frac{1}{N}\sum_{i=1}^{N} f_i(Z_1, \cdots, Z_i) > \exp\left(-\frac{n\epsilon}{2}\right)\right\} < \exp\left[-\frac{N}{2}\exp\left(-\frac{n\epsilon}{2}\right)\right] < \exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{2}\right)\right]$$

where in the last step we used the assumption $N \ge \exp(n\epsilon)$.

Hence by (A11),

$$\Pr\left\{\frac{1}{N}\left|\{i: Z_i \in \tau_{X|X'S}(Z_j, s) \text{ for some } j < i\}\right|\right\} < \exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{2}\right)\right] + \exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{4}\right)\right] < 2\exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{4}\right)\right].$$

By symmetry, the same holds when "for some $j < i$" is replaced by "for some $j > i$." Thus we finally obtain that

$$\Pr\left\{\frac{1}{N}\left|\{i: Z_i \in \tau_{X|X'S}(Z_j, s) \text{ for some } j \ne i\}\right|\right\} < 4\exp\left[-\frac{1}{2}\exp\left(\frac{n\epsilon}{4}\right)\right] \quad \text{(A12)}$$

if $I(X \wedge X'S) > |R - I(X' \wedge S)|^+ + \epsilon$ and $n \ge n_1(\epsilon/4)$.

Now the proof is completed in the standard manner. As the total number of all possible combinations of sequences $x \in \tau_X$, $s \in \mathcal{S}^n$ and joint types $P_{XX'S}$ grows only exponentially with n, the doubly exponential probability bounds (A7), (A9), and (A12) ensure that with probability close to 1 all the inequalities

$$\left|\{j: z_j \in \tau_{X'|XS}(x, s)\}\right| \le \exp\left[n\left(|R - I(X' \wedge XS)|^+ + \epsilon\right)\right]$$

$$\frac{1}{N}\left|\{j: z_j \in \tau_{X'|S}(s)\}\right| \le \exp\left(-\frac{n\epsilon}{2}\right) \quad \text{if } I(X' \wedge S) \ge \epsilon$$

and

$$\frac{1}{N}\left|\{i: z_i \in \tau_{X|X'S}(z_j, s) \text{ for some } j \ne i\}\right| \le \exp\left(-\frac{n\epsilon}{2}\right), \quad \text{if } I(X \wedge X'S) > |R - I(X' \wedge S)|^+ + \epsilon,$$

hold simultaneously if n is sufficiently large, $n \ge n_0(\epsilon)$. Any realization of the random N-tuple $\{Z_1, \cdots, Z_N\}$ simultaneously satisfying all these inequalities is a proper choice for $\{x_1, \cdots, x_N\}$ in Lemma 2.

Lemma A2: For a nonsymmetrizable AVC, there exists $\xi > 0$ such that for each pair of channels $U_1 = \mathcal{X} \to \mathcal{S}$, $U_2 = \mathcal{X} \to \mathcal{S}$,

$$\max_{x, x', y}\left|\sum_s W(y|x,s)U_1(s|x') - \sum_s W(y|x',s)U_2(x|s)\right| \ge \xi. \quad \text{(A13)}$$

Further, for any AVC and $\alpha > 0$, there exists $\xi > 0$ such that (A13) holds for every U_1 and U_2 for which a P can be found with

$$\sum_{x,s} P(x)U_1(s|x)l(x) \le \Lambda_0(P) - \alpha$$

$$\sum_{x,s} P(x)U_2(s|x)l(x) \le \Lambda_0(P) - \alpha. \quad \text{(A14)}$$

Proof: The maximum in (A13) does not change upon interchanging the two sums and then x and x'. Thus

$$\begin{aligned} &\max_{x, x'y}\left|\sum_s W(y|x,s)U_1(s|x') - \sum_s W(y|x',s)U_2(s|x)\right| \\ &= \max_{x, x', y}\left|\sum_s W(y|x,s)U_2(s|x') - \sum_s W(y|x',s)U_1(s|x)\right| \\ &\ge \max_{x, x', y}\left|\sum_s W(y|x,s)U(s|x') - \sum_s W(y|x',s)U(s|x)\right| \end{aligned} \quad \text{(A15)}$$

where $U = (1/2)(U_1 + U_2)$. Notice that if U_1 and U_2 satisfy (A14) for some P, then it holds that

$$\sum_{x,s} P(x)U(s|x)l(s) \le \Lambda_0(P) - \alpha. \tag{A16}$$

Denote the last maximum in (A15) by $F(U)$. As a continuous function on the compact set of all channels $U: \mathcal{X} \to \mathcal{S}$, $F(U)$ attains its minimum at some U^*. If the AVC is nonsymmetrizable, U^* cannot satisfy (2.7), and hence $F(U^*) > 0$. This proves the first assertion with $\xi = F(U^*)$. Further, by considering $F(U)$ as a continuous function of (P, U) ranging over the compact set of all pairs (P, U) satisfying (A16), we see that its minimum is attained for some (P^*, U^*). As (P^*, U^*) satisfies (A16), U^* cannot satisfy (2.7) (by (2.13)). Hence once again $F(U^*) > 0$, completing the proof of Lemma A2.

References

[1] R. Ahlswede, "A note on the existence of the weak capacity for channels with arbitrarily varying channel probability functions and its relation to Shannon's zero error capacity," *Ann. Math. Statist.*, vol. 41, p. 1027, 1970.

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

[3] 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.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[5] ____, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 57, pp. 87–101, 1981.

[6] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," preprint, presented at the IEEE Int. Symp. Information Theory, Cornell University, Ithaca, NY, 1977.

[7] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 27–34, Jan. 1988.

[8] R. L. Dobrushin and S. Z. Stambler, "Coding theorems for classes of arbitrarily varying discrete memoryless channels," *Probl. Peredach. Inform.*, vol. 11, no. 2, pp. 3–22, 1975 (English translation).

[9] 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.

[10] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964, English translation.

[11] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. 2, pp. 8–19, 1956.

[12] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. Berlin: Springer-Verlag, 1978.