Capacity and Decoding Rules for Classes of Arbitrarily Varying Channels

IMRE CSISZÁR AND PRAKASH NARAYAN, MEMBER, IEEE

Abstract —We consider the capacity of an arbitrarily varying channel (AVC) for deterministic codes with the average probability of error criterion and, typically, subject to a state constraint. First, sufficient conditions are provided that enable (relatively) simple decoding rules such as typicality, maximum mutual information, and minimum distance, to attain capacity. Then the (possibly noisy) OR channels and group adder channels are studied in detail. For the former the capacity is explicitly determined and shown to be attainable by minimum distance decoding. Next, for a large class of additive AVC's, in addition to providing an intuitively suggestive simplification of the general AVC capacity formula, we prove that capacity can be attained by a universal decoding rule. Finally, the effect of random state selection on capacity is studied, enabling us to comment on the merits and limitations of a previous "mutual information game" approach.

I. Introduction

THE CAPACITY of an *arbitrarily varying channel* (AVC) for deterministic codes and the average probability of error criterion is, in the absence of channel state constraints, either equal to its capacity for random codes or else to zero (Ahlswede's alternatives [1]). In an earlier paper [9] we have shown that the latter contingency arises only for the trivial reason of symmetrizability, and determined capacity also in the presence of a state constraint when Ahlswede's alternatives have been shown no longer to hold.

This paper is a continuation of [9]. As in [9], we consider discrete memoryless AVC's and deterministic codes with the average probability of error criterion. The communication situation modeled is one wherein both encoder and decoder are ignorant of the channel state sequence, and the state selector—though cognizant of the code—is ignorant of the message transmitted. While a knowledge of the basic terminology and notation of [9] is assumed of the reader, further familiarity with [9] is not necessary to understand this paper, with the sole exception of a proof in Appendix II.

In Section II the problem of decoding for AVC's is addressed. The decoding rule used in [9] was quite complex. It did not belong to the class of α-decoding rules in the sense of Csiszár–Körner [5], i.e., it could not be defined solely in terms of the joint types of the codeword–received sequence pairs; in fact, triple joint types were also involved. A need for complex decoders for AVC's also arose with the maximum probability of error criterion (cf. Ahlswede [2] and Csiszár–Körner [6]). To our knowledge the first nonstandard decoding rule in Shannon theory appears in [2]. While finding a good decoder in the class of α-decoding rules for every AVC appears unlikely, a sufficient condition set forth in Theorem 1 and its corollary nevertheless will enable us to demonstrate that certain common α-decoding rules suffice for the classes of AVC's considered in later sections. In particular, Theorems 2 and 3, respectively, provide sufficient conditions for the efficacy of the maximum mutual information (MMI) and typically decoding rules. The former possesses the desirable feature of being universal, i.e., independent of the given channel. The capacity of a discrete memoryless channel can always be attained by using the MMI decoding rule (Goppa [11], Csiszár–Körner [7, sec. 2.5]). Decoding by joint typicality has been employed by Dobrushin–Stambler [10]; Theorem 3 is closely related to their result. For the class of additive AVCs, the typicality decoding rule is partically equivalent to a universal one which we term the independence decoding rule. A sufficient condition for its appropriateness for a given AVC is provided in Theorem 4. For AVC's with binary input and output alphabets, Theorem 5 identifies a condition for the simple minimum distance decoding rule to be effective. This condition will apply, in particular, to the OR channel considered in Section III. For an application of minimum distance decoding to binary AVC's with the maximum probability of error criterion, see Ahlswede–Wolfowitz [4].

In Sections III and IV some interesting examples of AVC's are considered. In Section III we determine the capacity under a state constraint of the OR channel with and without noise, and comment on the solution of a combinatorial problem as a special case of our results for the noiseless OR channel. Section IV deals with the class of AVC's whose inputs, states, and outputs belong to (finite subsets of) a possibly infinite commutative group $\mathcal{G}$, with the output being determined by the group addition of input, state, and possibly noise. The binary and arithmetic

Manuscript received November 23, 1987; revised October 11, 1988. This work was supported in part by the Minta Marta Fund for Aerospace Research and by the Systems Research Center at the University of Maryland under the National Science Foundation Engineering Research Center Grant OIR-85-00108. This paper was presented in part at the IEEE International Symposium on Information Theory, Kobe, Japan, June 19–24, 1988.

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 and the Systems Research Center, University of Maryland, College Park, MD 20742, USA.

IEEE Log Number 8929033.

0018-9448/89/0700-0752$01.00 ©1989 IEEE

adder AVC's of [9] are simple (noiseless) cases of such group adder AVC's. For finite $\mathcal{G}$ with the input alphabet being equal to $\mathcal{G}$, Theorem 7 provides a simple formula for the capacity under a given state constraint without requiring any assumptions on the particular cost function appearing therein. For the general case we also obtain a useful, albeit less explicit, result (Theorem 8).

In Section V we consider additive AVC's with vector addition as the group operation but otherwise more general than the group adder AVC's of Section IV; in particular, the noise is also allowed to be arbitrarily varying. Under some hypotheses on the input and state constraints, Theorem 9 establishes the intuitive result that the capacity is positive if and only if the state constraint is more restrictive than the input constraint and shows that capacity can be attained by the independence decoding rule. Also, for these AVC's the general capacity formula of [9] is somewhat simplified.

The effect of various kinds of randomized state selection on capacity is considered in Section VI. Theorem 10 determines the capacity for three different versions of independent state selection subject to an expectation constraint. For dealing with channels partially controlled by an adversary, McEliece [14] has suggested a game-theoretic approach with mutual information as the pay-off function. The results of Section VI will enable us to specify the conditions under which this approach is justified from the viewpoint of AVC theory.

Finally, we shall show in Appendix I that the previous sufficient conditions for positive capacity due to Dobrushin–Stambler [10] and Ahlswede [1] are not necessary in general. Ahlswede [1] had also obtained a necessary and sufficient condition which, in the terminology of multiuser Shannon theory, was a noncomputable "product space characterization" (cf. [7, p. 259]); we shall indicate that nonsymmetrizability may be regarded as a "single letterization" of that condition. We now recall the main results of [9] which will be used throughout this paper.

Given an AVC W with input alphabet $\mathcal{X}$, set of states $\mathcal{S}$, and output alphabet $\mathcal{Y}$, let us denote by W_Q, for any distribution Q on $\mathcal{S}$, the channel $\mathcal{X} \to \mathcal{Y}$ defined by

$$W_Q(y|x) = \sum_{s \in S} W(y|x,s)Q(s). \tag{1.1}$$

The mutual information $I(X \wedge Y)$ between random variables with joint distribution $P_{XY}(x,y) = P(x)W_Q(y|x)$ will be denoted by $I(P, W_Q)$.

As in [9], we denote by $\mathcal{U}$ the set of channels U: $\mathcal{X} \to \mathcal{S}$ such that for every $x \in \mathcal{X}$, $x' \in \mathcal{X}$, $y \in \mathcal{Y}$

$$\sum_{s \in \mathcal{S}} W(y|x,s)U(s|x') = \sum_{s \in \mathcal{S}} W(y|x',s)U(s|x). \tag{1.2}$$

The AVC is said to be *symmetrizable* if and only if $\mathcal{U} \neq \phi$. By Theorem 1 of [7], the capacity C of an (unconstrained) AVC is zero if and only if the AVC is symmetrizable; for a nonsymmetrizable AVC

$$C = \max_P \min_Q I(P, W_Q). \tag{1.3}$$

The simplest symmetrizable AVC's are those that are symmetric, i.e., $\mathcal{X} = \mathcal{S}$, and $W(y|x,s) = W(y|s,x)$ for every x and s. In this case, (1.2) holds with U being the identity matrix. Another simple instance of symmetrizability, termed deterministic symmetrizability, arises when (1.2) holds for some deterministic channel U, i.e., with a matrix whose entries are $\{0,1\}$-valued. We remark, however, that even for deterministic AVC's with $\mathcal{X} = \mathcal{S}$, deterministic symmetrizability is not a necessary condition for $C = 0$ (cf. Example 1 in Appendix I).

We also recall the concept of a state constraint Λ, which permits only those state sequences $s = (s_1, \cdots, s_n)$ that satisfy

$$l(s) = \frac{1}{n} \sum_{i=1}^{n} l(s_i) \leq \Lambda \tag{1.4}$$

where l is a given nonnegative-valued function on $\mathcal{S}$ with $\min_s l(s) = 0$.

The capacity under state constraint Λ may be positive even if the AVC is symmetrizable, and depends on the functional

$$\Lambda_0(P) = \min_{U \in \mathcal{U}} \sum_{x,s} P(x)U(s|x)l(s). \tag{1.5}$$

Thus by the corollary of Theorem 3 of [9], the capacity under state constraint Λ is zero if Λ is greater than

$$\Lambda_0 = \max_P \Lambda_0(P) = \min_{U \in \mathcal{U}} \max_{x \in \mathcal{X}} \sum_s U(s|x)l(s) \tag{1.6}$$

(with $\Lambda_0(P) = \infty$ if $\mathcal{U} = \phi$), while it is positive and equals

$$C(\Lambda) = \max_{P:\ \Lambda_0(P) \geq \Lambda} I(P, \Lambda) \quad \text{if} \quad \Lambda < \Lambda_0 \tag{1.7}$$

where

$$I(P, \Lambda) = \min_{Q:\ l(Q) \leq \Lambda} I(P, W_Q) \tag{1.8}$$

with

$$l(Q) = \sum_{s \in \mathcal{S}} Q(s)l(s). \tag{1.9}$$

In particular, it is possible that $C(\Lambda)$ lies strictly between 0 and the random code capacity under state constraint Λ (cf. [8]), which equals

$$C_r(\Lambda) = \max_P I(P, \Lambda) \tag{1.10}$$

with no constraint on the distribution P.

Remark: In [9], Λ_0 was defined as $\max_P \Lambda_0(P)$. The alternative expression in (1.6) results by observing that from (1.5) and the minimax theorem

$$\max_P \Lambda_0(P) = \min_{U \in \mathcal{U}} \max_P \sum_{x,s} P(x)U(s|x)l(s)$$

and further, that the inner maximum is attained when P is concentrated at a point $x \in \mathcal{X}$ maximizing $\Sigma_s U(s|x)l(s)$.

In [9], we had in fact determined the capacity $C(\Gamma, \Lambda)$ for the case where the codewords $x = (x_1, \cdots, x_n)$ were required to satisfy an input constraint

$$g(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \leq \Gamma \tag{1.11}$$

for some given function g on $\mathscr{X}$ and constant Γ. Then the capacity with no input constraint, denoted by $C(\Lambda)$, was obtained as $C(g_{\max}, \Lambda)$ because the constraint (1.11) became inactive if $\Gamma = g_{\max}$. By [9, theorem 3], with $g(P)$ defined as in [9, eq. (1.9)],

$$C(\Gamma, \Lambda) = \begin{cases} \max\limits_{\substack{P:\, \Lambda_0(P) \geq \Lambda \\ g(P) \leq \Gamma}} I(P, \Lambda) > 0, & \text{if } \max\limits_{P:\, g(P) \leq \Gamma} \Lambda_0(P) > \Lambda \\ 0, & \text{if } \max\limits_{P:\, g(P) \leq \Gamma} \Lambda_0(P) < \Lambda. \end{cases} \tag{1.12}$$

In the case $\max_{P:\, g(P) \leq \Gamma} \Lambda_0(P) = \Lambda$, left undecided by this theorem, we could still assert $C(\Gamma, \Lambda) = 0$ if (for every P) the minimum in (1.5) was attained by a $0-1$ matrix $U \in \mathscr{U}$ (cf. the remark following the proof of [9, theorem 3]). Again, it is possible that $C(\Gamma, \Lambda)$ lies strictly between 0 and the random code capacity under state constraint Λ and input constraint Γ (cf. [8]), which equals

$$C_r(\Gamma, \Lambda) = \max_{P:\, g(P) \leq \Gamma} I(P, \Lambda). \tag{1.13}$$

II. Decoding Rules for AVC's

We now address the problem of whether, and for which AVC's, simpler decoding rules than that employed in [9] are effective. In this section we shall always deal with AVC's with state constraint Λ. These, of course, include unconstrained AVC's as well by simply setting $\Lambda > \max_s l(s)$. To avoid repetition, we proceed with the understanding that an AVC with state constraint Λ is given without explicitly mentioning this in the definitions and theorems.

It will be easier to present our results in a mathematically satisfactory manner if we distinguish between a decoding rule and a decoder. Recall that a code of blocklength n, with message set $\{1, \cdots, N\}$, is a pair of mappings $f\colon \{1, \cdots, N\} \to \mathscr{X}^n$, $\phi\colon \mathscr{Y}^n \to \{0, \cdots, N\}$; here f is the encoder, ϕ is the decoder, and a decoder output 0 means that an error has been declared.

By a *decoding rule* we shall mean a prescription for defining a decoder when the codewords $\boldsymbol{x}_i = f(i)$, $i = 1, \cdots, N$, are given. For convenience, we shall permit this prescription to depend also on a parameter, typically, a threshold that can be "suitably" chosen. For some received sequences $\boldsymbol{y} \in \mathscr{Y}^n$, the decoder may assign more messages than one as "candidate" decoder outputs. In such a case, however, none of the candidates is accepted; instead, an error is declared (i.e., the decoder output is set equal to 0). Of course, an error is declared also if the rule assigns no "candidate" message to $\boldsymbol{y}$. This convention will facilitate the proof of the efficacy of specific decoding rules (for suitable classes of AVC's), such as the MMI and typicality decoding rules defined later.

The *maximum mutual information* (MMI) *decoding rule* is universal (i.e., it does not depend on the given AVC and state constraint). Given the codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$, it assigns to each $\boldsymbol{y} \in \mathscr{Y}^n$ the message i maximizing the nonprobabilistic mutual information $I(\boldsymbol{x}_i \wedge \boldsymbol{y})$ (in case of a tie, an error is declared). Here, for sequences $\boldsymbol{x} \in \mathscr{X}^n$, $\boldsymbol{y} \in \mathscr{Y}^n$, $I(\boldsymbol{x} \wedge \boldsymbol{y})$ is defined as the mutual information of dummy random variables representing the joint type of $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e.,

$$I(\boldsymbol{x} \wedge \boldsymbol{y}) = I(\tilde{X} \wedge \tilde{Y}), \qquad \text{where } P_{\tilde{X}\tilde{Y}} = P_{\boldsymbol{x}, \boldsymbol{y}}. \tag{2.1}$$

The *typicality decoding rule* assigns message i to a received sequence $\boldsymbol{y} \in \mathscr{Y}^n$ if, supposing that the codewords are of type P, the codeword $\boldsymbol{x}$ corresponding to message i satisfies

$$\max_{x, y} \left| P_{\boldsymbol{x}, \boldsymbol{y}}(x, y) - P(x) W_Q(y|x) \right| \leq \tau, \qquad \text{for some } Q \text{ with } l(Q) \leq \Lambda. \tag{2.2}$$

Here $\tau > 0$ is a constant and a parameter of the decoding rule that can be "suitably" chosen.

In the terminology of Csiszár–Körner [7, p. 34], condition (2.2) states that $\boldsymbol{y}$ is W_Q-typical under the condition $\boldsymbol{x}$, with constant τ, for some of the channels W_Q with $l(Q) \leq \Lambda$. For the purposes of this paper we will use the following abbreviated terminology for condition (2.2): $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-*typical.* Further, we will say that $\boldsymbol{y}$ is τ-typical if it is $(\boldsymbol{x}, \tau)$-typical in this sense for at least one codeword $\boldsymbol{x}$. Note that typicality in this special sense is defined only for received sequences and only if the codeword set is given. At this point, we recall our implicit assumption in this section that in all definitions and theorems an AVC with state constraint Λ is given.

Definition 1: A decoding rule will be called *good* for input distribution P if for any $\delta > 0$ a neighborhood of P can be found such that for sufficiently large block lengths n, for any type P' in this neighborhood, there exist codes with codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ of type P' and decoder specified by the given decoding rule (possibly with a parameter depending on δ) with rate

$$\frac{1}{n} \log N > I(P, \Lambda) - \delta \tag{2.3}$$

and average probability of error

$$\bar{e}(\boldsymbol{s}) \leq \delta, \qquad \text{for every } \boldsymbol{s} \in \mathscr{S}^n \text{ with } l(\boldsymbol{s}) \leq \Lambda. \tag{2.4}$$

Further, we say that capacity can be attained by the given decoding rule if this rule is good for input distributions P such that $I(P, \Lambda)$ is arbitrarily close to $C(\Lambda)$.

The next definition will enable us to formulate concisely a useful sufficient condition for the goodness of a decoding rule, viz., Theorem 1 and its corollary.

Definition 2: A decoding rule will be called (ξ, τ)-*admissible* for codeword type P if it assigns to each τ-typical $\boldsymbol{y} \in \mathscr{Y}^n$ at least one candidate message i, and for $\boldsymbol{x}' = \boldsymbol{x}_i$,

a) $I(\boldsymbol{x}' \wedge \boldsymbol{y}) \geq I(P, \Lambda) - \xi$,
b) for each codeword $\boldsymbol{x}$ such that $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-typical, letting X, X', Y be dummy random variables representing the joint type of $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}$, and for an arbitrary

S with distribution Q (say), such that

$$\max_{x,s,y} |P_{XSY}(x,s,y) - P(x)Q(s)W(y|x,s)| \le \tau, \qquad l(Q) \le \Lambda \quad (2.5)$$

we have

$$I(XY \wedge X'|S) > \tau. \quad (2.6)$$

Remark: When a decoding rule involves a parameter, we will permit the latter to be selected depending on ξ and τ in Definition 2. In particular, the parameter τ of the joint typicality decoding rule will be set equal to the τ of Definition 2.

Theorem 1: For any $\xi > 0$, there exist $\tau > 0$ and $\gamma > 0$ such that for sufficiently large blocklength n, for any type P, there exist codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ of type P with

$$\frac{1}{n}\log N > I(P, \Lambda) - 3\xi$$

such that if the decoder is defined by a (ξ, τ)-admissible decoding rule, then

$$\max_{\boldsymbol{s}:\, l(\boldsymbol{s}) \le \Lambda} \bar{e}(\boldsymbol{s}) \le \exp(-n\gamma).$$

Corollary: A decoding rule is good for an input distribution P if for every $\xi > 0$, there is a $\tau > 0$ and a neighborhood of P such that for codeword types in this neighborhood the given decoding rule is (ξ, τ)-admissible.

Proof: The proof is effectively contained in that of the main result of [9], which relied in essence of the admissibility of the decoding rule therein. For details, see Appendix II. The corollary immediately follows by the continuity of $I(P, \Lambda)$ as a function of P.

Theorem 2: For every $\alpha > 0$ and $\xi > 0$, there exists $\tau > 0$ such that the MMI decoding rule is (ξ, τ)-admissible for every codeword type P such that

$$I(X \wedge Y) \ge I(S \wedge Y) + \alpha \quad (2.7)$$

whenever $P_{XSY}(x,s,y) = P(x)Q(s)W(y|x,s)$ with $l(Q) \le \Lambda$. In particular, the MMI decoder is good for an input distribution P if $I(X \wedge Y) > I(S \wedge Y)$ for P_{XSY} as before.

Proof: Let $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}$, and the dummy random variables X, X', Y, S be as in Definition 2. Then by the definition of the MMI decoding rule, $I(\boldsymbol{x}' \wedge \boldsymbol{y}) = \max_j I(\boldsymbol{x}_j \wedge \boldsymbol{y})$, and in particular,

$$I(X' \wedge Y) = I(\boldsymbol{x}' \wedge \boldsymbol{y}) \ge I(\boldsymbol{x} \wedge \boldsymbol{y}) = I(X \wedge Y). \quad (2.8)$$

As $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-typical, we obtain from (2.1) and (2.2) that $I(\boldsymbol{x} \wedge \boldsymbol{y})$ is close to $I(P, W_Q) \ge I(P, \Lambda)$ if τ is sufficiently small. This and (2.8) establish a) in Definition 2. Turning to b), observe that

$$\begin{aligned} I(XY \wedge X'|S) &\ge I(Y \wedge X'|S) = I(Y \wedge X'S) - I(Y \wedge S) \\ &\ge I(Y \wedge X') - I(Y \wedge S) \\ &\ge I(X \wedge Y) - I(Y \wedge S), \end{aligned}$$

where the last step follows from (2.8). Since X, S, Y satisfies (2.5), i.e., their joint distribution is arbitrarily close to a joint distribution for which the condition (2.7) has been postulated, the required inequality (2.6) follows by continuity, if τ is sufficiently small. This completes the proof of the first assertion of Theorem 2.

Finally, if for some fixed P we have $I(X \wedge Y) > I(S \wedge Y)$ whenever $P_{XSY}(x,s,y) = P(x)Q(s)W(y|x,s)$ with $l(Q) \le \Lambda$, then by continuity, for a sufficiently small $\alpha > 0$, (2.7) will hold for joint distributions of the latter kind even if P is replaced by P', in a sufficiently small neighborhood of P. Then the last assertion of Theorem 1 follows from the first one and the corollary of Theorem 1.

To obtain an analogous result for the typicality decoding rule, we define the following conditions, the first of which is effectively due to Dobrushin–Stambler (DS) [10].

Definition 3: A distribution P on $\mathscr{X}$ is said to satisfy Condition DS if no distribution Q on $\mathscr{S}$ and channel U: $\mathscr{X} \to \mathscr{S}$ exist such that for every $x' \in \mathscr{X}$ and $y \in \mathscr{Y}$

$$\sum_{x,s} P(x)W(y|x,s)U(s|x') = \sum_{s} W(y|x',s)Q(s). \quad (2.9)$$

Further, P is said to satisfy (the weaker) Condition DS (Λ) if no Q and U exist satisfying (2.9) and, in addition (cf. (1.9))

$$l(Q) \le \Lambda, \qquad \sum_{s} P(x)U(s|x)l(s) \le \Lambda. \quad (2.10)$$

Condition DS states that by putting any channel U: $\mathscr{X} \to \mathscr{S}$ in cascade with the channel W^P: $\mathscr{S} \to \mathscr{Y}$ defined by

$$W^P(y|s) = \sum_{x} P(x)W(y|x,s),$$

the resulting channel UW^P: $\mathscr{X} \to \mathscr{Y}$ cannot be of the form W_Q as in (1.1). A better understanding of this condition, which is not too perspicuous, may be obtained by comparing it with other relevant conditions for AVC's (cf. Appendix I).

Theorem 3: The typicality decoding rule is good for every strictly positive P which satisfies Condition DS (Λ), i.e., for which (2.9) and (2.10) cannot simultaneously hold for any Q and U.

Corollary: Capacity can be attained by the typicality decoding rule if some input distribution P with $I(P, \Lambda) = C(\Lambda)$ satisfies DS(Λ). Here the strict positivity of P is not required.

Proof: This is an easy consequence of the corollary of Theorem 1. The role of the hypothesis on P is to ensure the validity of condition (b) in Definition 2. To see this heuristically, let X, X', S, Y be as in Definition 2, and assume that the codeword type actually equals P, and $\tau = 0$. Then

$$\begin{aligned} P_{X'S}(x',s) &= P(x')W_Q(y|x') \\ P_{XSY}(x,s,y) &= P(x)\tilde{Q}(s)W(y|x,s) \quad (2.11) \end{aligned}$$

for some Q and $\tilde{Q}$ with $l(Q) \le \Lambda$, $l(\tilde{Q}) \le \Lambda$, and condition b) requires that $I(XY \wedge X'|S)$ be nonzero. However,

$I(XY \wedge X'|S) = 0$ means that

$$P_{XX'SY}(x, x', s, y) = P_{X'S}(x', s) P_{XY|S}(x, y|s).$$

Upon dividing both sides by $P(x')$ (permissible by the strict positivity assumption) and summing over x and s, we get

$$P_{Y|X'}(y|x') = \sum_{s \in \mathcal{S}} P_{S|X'}(s|x') P_{Y|S}(y|s). \qquad (2.12)$$

This, however, contradicts the hypothesis on P as (2.9) and (2.10) would now hold for P, Q, and $U = P_{S|X'}$.

Since τ, though arbitrarily small, cannot be set equal to 0, a rigorous proof entails some technical details and is, therefore, deferred to Appendix II.

To see that the strict positivity of P is not needed in the corollary, observe that the set of P's for which DS (Λ) holds is a closed set. Hence its complement is an open set and, therefore, any P in the corollary can be approximated by strict positive distributions to which Theorem III can be applied.

Remarks: 1) For the special case of an unconstrained AVC, Theorem 3 states that the typicality decoder is good for every strictly positive P which satisfies DS. This assertion is almost identical to the main result of Dobrushin–Stambler [10], whose failure to completely solve the AVC capacity problem stemmed from their reliance on the typicality decoding rule. In particular, Theorem 3 provides a direct proof of the fact that an unconstrained AVC, which has some input distribution P satisfying Condition DS, or an AVC with state constraint Λ which has some input distribution P satisfying Condition DS (Λ), has positive capacity.

2) Stambler's [15] theorem on the capacity of the AVC *with states known at the receiver* can also be obtained from Theorem 3, applying it to the AVC $\tilde{W}$ whose output alphabet is $\mathcal{Y} \times \mathcal{S}$ and $\tilde{W}(y, s'|x, s) = W(y|x, s)$ if $s' = s$, and 0, otherwise (cf. [7], pp. 227). Indeed, for $\tilde{W}$ every input distribution P satisfies Condition DS, except for the trivial case where for some $s \in \mathcal{S}$ the channel $W(\cdot|\cdot, s)$ has capacity zero. To see this, notice that for $\tilde{W}$, (2.9) becomes

$$\sum_x P(x) W(y|x, s) U(s|x') = W(y|x', s) Q(s).$$

This, however, implies that $W(y|x', s)$ is independent of x' whenever $Q(s) \neq 0$ because summing over y gives that $U(s|x') = Q(s)$.

For the important class of *additive* ADC's the typicality decoding rule is practically equivalent to what we term the *independence decoding rule*, which has the merit of being universal. In fact, for additive AVC's the hypothesis of Theorem 3 will also imply the goodness of the latter decoding rule; this will be established in Theorem 4 to follow.

An AVC W will be called *additive* if $W(y|x, s)$ depends on x and y through the difference $y - x$ only. Of course, this requires that $\mathcal{X}$ and $\mathcal{Y}$ be subsets of a commutative group $\mathcal{G}$. Here $\mathcal{G}$ need not be finite and, indeed, in Section 5 we will consider additive AVCs with $\mathcal{G} = \mathbb{R}^d$.

Formally, an additive AVC with (finite) input alphabet $\mathcal{X} \subset \mathcal{G}$ and (finite) set of states $\mathcal{S}$ is defined by a channel V: $\mathcal{S} \to \mathcal{Z}$, where $\mathcal{Z}$ is a (finite) subset of $\mathcal{G}$, by setting

$$W(y|x, s) = V(y - x|s). \qquad (2.13)$$

It is understood here that $\mathcal{Y} = \{y\colon\ y = x + z,\ x \in \mathcal{X},\ z \in \mathcal{Z}\}$, and that $V(z|s) = 0$ for $z \notin \mathcal{Z}$.

For an additive AVC the independence decoding rule assigns, by definition, message i to a received sequence $\boldsymbol{y}$ whenever the codeword $\boldsymbol{x} = \boldsymbol{x}_i$ is *η-independent* of the error vector $\boldsymbol{y} - \boldsymbol{x}$ in the sense that

$$I(\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x}) \leq \eta. \qquad (2.14)$$

Here $\eta > 0$ is a parameter that can be suitably chosen.

To precisely formulate the claimed equivalence of the typicality and independence decoding rules for additive AVC's, viz. Lemma 1 to follow, we need a concept of regularity for probability distributions on $\mathcal{G}$. For distributions P and Q on $\mathcal{G}$ with finite support, the *convolution* $P * Q$ is defined by

$$(P * Q)(y) = \sum_{x \in \mathcal{X}} P(x) Q(y - x). \qquad (2.15)$$

We will say that P is *regular* if

$$P * Q_1 = P * Q_2 \text{ implies } Q_1 = Q_2. \qquad (2.16)$$

Observe that for the case $\mathcal{G} = \mathbb{R}^d$ this is automatically satisfied as the characteristic function of P (having a finite support) cannot vanish in an interval. If $\mathcal{G} = \{0, \cdots, k-1\}$ with mod-k addition, then (2.16) holds if and only if the polynomial $p(D) = \sum_{i=0}^{k-1} P(i) D^i$ is not a divisor of $D^k - 1$. In particular, for $k = 2$ or 3, only the uniform distribution on $\mathcal{G}$ is nonregular. For $k \geq 4$, there are others, but only finitely many.

Lemma 1: Let an additive AVC be given with state constraint Λ. Then

1) for any $\eta > 0$ there exists $\tau > 0$ such that if $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-typical for some codeword $\boldsymbol{x}$, then $\boldsymbol{x}$ and $\boldsymbol{y} - \boldsymbol{x}$ are η-independent;

2) for any closed set $\mathcal{P}$ of distributions on $\mathcal{X}$ which are regular in the sense of (2.16), and for any $\tau > 0$, there exists $\eta > 0$ such that if $\boldsymbol{x}$ and $\boldsymbol{y} - \boldsymbol{x}$ are η-independent, then $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-typical provided that the codewords are of type $P \in \mathcal{P}$ and that $\boldsymbol{y}$ is η-typical.

Notice that if $\mathcal{G} = \mathbb{R}^d$ then the restriction in part 2) of this lemma is immaterial and $\mathcal{P}$ can be taken to be the set of all distributions on $\mathcal{X}$.

Proof: See Appendix II.

Theorem 4: For an additive AVC the independence decoding rule is good for every regular input distribution P which is strictly positive and satisfies Condition DS (Λ).

Corollary: Under the hypothesis of the corollary of Theorem 3 capacity can be attained by the independence decoding rule.

Proof: By Definition 2 and Lemma 1, part 2), if the typicality decoding rule is (ξ, τ)-admissible for codeword types in $\mathscr{P}$ (with any $\xi > 0$), then the independence decoding rule is (ξ, η)-admissible for $\mathscr{P}$, τ, and η as in Lemma 1, part 2. Hence the proof of Theorem 3 establishes Theorem 4 as well if we observe that all distributions in a sufficiently small neighborhood of a regular P are also regular.

The corollary follows as before. In the corollary, the regularity hypothesis need not be imposed as any distribution can be approximated by a sequence of regular (and strictly positive) ones.

For channels with binary inputs and outputs, the simple *minimum distance* (MD) *decoding rule* often suffices. This rule assigns to a received sequence y the message $i \in \{1, \cdots, N\}$ that minimizes the Hamming distance $d(x_i, y)$ (in case of ties an error is declared). We conclude this section by determining a general relation between MD and MMI. This will enable us to establish the efficacy of the MD decoding rule for important classes of AVC's. We need the following simple lemma.

Lemma 2: Let X, X', Y be binary random variables such that 1) X and X' have the same distribution $(1-p, p)$, and 2) in the conditional distribution matrix of Y given X, the sum of the minor diagonal elements does not exceed 1. Then

$$\Pr\{X' \neq Y\} \le \Pr\{X \neq Y\} \text{ implies } I(X' \wedge Y) \ge I(X \wedge Y).$$

Proof: Fixing Y, consider the class of all joint distributions P_{XY} with X-marginal $(1-p, p)$. Clearly, a joint distribution in this class is uniquely determined by the probability $\Pr\{X \neq Y\}$. In particular, the conditional probability matrix V of Y given X can be represented as a function of $\Pr\{X \neq Y\}$; moreover, this function is linear. Since $I(X \wedge Y)$ is a convex function of the conditional probability matrix V, it follows that (within the considered class of joint distributions) it is a convex function also of $\Pr\{X \neq Y\}$. This function attains its minimum when X and Y are independent, i.e., when $V(1|0) + V(0|1) = 1$, and, therefore, it is monotone decreasing for those values of the argument $\Pr\{X \neq Y\}$ that correspond to a conditional probability matrix with $V(1|0) + V(0|1) \le 1$ (since the latter sum is an increasing function of $\Pr\{X \neq Y\}$).

Theorem 5: For an AVC with binary input and output alphabets such that

$$W_Q(1|0) + W_Q(0|1) < 1 \qquad \text{whenever } l(Q) \le \Lambda \quad (2.17)$$

Theorem 2 remains true when "MMI" is replaced by "MD."

Proof: Given the codewords, all of type P, and a τ-typical received sequence y, let x be any codeword for which y is (x, τ)-typical, and let x' have minimum Hamming distance from y. Let X, X', Y be dummy random variables representing the joint type of x, x', y. The proof of Theorem 2 will then apply *verbatim* if we establish (2.8).

To this end, observe that

$$\Pr\{X' \neq Y\} = \frac{1}{n} d(x', y) \le \frac{1}{n} d(x, y) = \Pr\{X \neq Y\}.$$

By Lemma 2, this implies (2.8) provided that condition 2) of that Lemma is satisfied. However, this is ensured by hypothesis (2.17) if τ is sufficiently small, because $P_{XY} = P_{x,y}$ satisfies (2.2). This completes the proof.

Remark: Dobrushin–Stambler [10] provided the following example. Let an AVC with $\mathscr{X} = \mathscr{Y} = \{0,1\}$, $\mathscr{S} = \{0,1,2,3\}$ be determined by

$$\begin{array}{lll} W(0|0,0) = 1 & W(0|0,1) = 0.6 & W(0|0,2) = 0.6 \\ W(0|0,3) = 1 & W(0|1,0) = 0.4 & W(0|1,1) = 0 \\ W(0|1,2) = 0.4 & W(0|1,3) = 0. & \end{array}$$

The capacity of this AVC is positive and is attained by $P = (0.5, 0.5)$. However, as this P does not satisfy Condition DS, the capacity of this AVC could not be determined by the approach of [10]. Ahlswede [1] found this fact somewhat disappointing, particularly as this channel was covered by earlier results of Ahlswede–Wolfowitz [4] which involved less mathematical effort. It is reassuring that Theorem 5 does apply to this example, guaranteeing that capacity can be attained by MD decoding (we omit the straightforward but none too enlightening calculations).

III. The OR Channel

The OR channel without noise is the AVC with $\mathscr{X} = \mathscr{S} = \mathscr{Y} = \{0,1\}$, whose output is obtained from its input and state by the Boolean operation OR. We also permit additive binary $(1-r, r)$ noise with a known $r < 1/2$. The *noisy* OR channel is determined by

$$\begin{aligned} W(0|0,0) = 1 - r \qquad & W(0|0,1) = W(0|1,0) \\ & = W(0|1,1) = r \end{aligned} \qquad (3.1)$$

and $W(1|x,s) = 1 - W(0|x,s)$. The *noiseless* OR channel is formally defined by (3.1) with $r = 0$.

The OR channel may model a multiple-access channel that performs, from the point of view of a single user, the OR operation on bits transmitted concurrently by other users. It may also be a model of a computer memory with defects, with state 1 representing a defective cell stuck at 1. Commencing with the work of Kuznetsov and Tsybakov [13], coding for such channels has been extensively studied (cf. Heegard and El Gamal [12], and references therein). Whereas in [12] the performance criterion involves averaging over an ensemble of state sequences, our AVC approach in contrast requires the error probability to be uniformly small over all feasible state sequences (however, unlike in [12], side information about the states is not available to the sender or the receiver).

Obviously, the OR channel defined by (3.1) has, without any state constraint, capacity zero (even the random code capacity is zero). We will adopt the constraint that the fraction of occurrences of $s = 1$ does not exceed a given $\Lambda < 1$. This comprises a "state constraint Λ" in the sense of (1.4) with $l(s) = s$.

To determine the capacity $C(\Lambda)$, first observe that for $Q = (1-q, q)$, the channel W_Q: $\mathscr{X} \to \mathscr{Y}$ (cf. (1.1)) is determined by

$$W_Q(0|0) = qr + (1-q)(1-r) = r + (1-q)(1-2r),$$
$$W_Q(0|1) = r. \tag{3.2}$$

Hence for $P = (1-p, p)$, the mutual information $I(P, W_Q)$ denoted by $I(p,q)$, equals

$$I(p,q) = h(r + (1-p)(1-q)(1-2r)) - (1-p)h(r + (1-q)(1-2r)) - ph(r) \tag{3.3}$$

where $h(t) = -t\log t - (1-t)\log(1-t)$. It is intuitively obvious, and easily verified by differentiation, that $I(p,q)$ is a decreasing function of q. With (1.9) yielding $l(Q) = q$, it follows from (1.8) that

$$I(P,\Lambda) = \min_{q \le \Lambda} I(p,q) = I(p,\Lambda). \tag{3.4}$$

Next, we determine $\Lambda_0(P)$, namely (cf. (1.5)) the minimum of

$$\sum_{x,s} P(x)U(s|x)l(s) = (1-p)U(1|0) + pU(1|1)$$

over the set of channels U: $\{0,1\} \to \{0,1\}$ that satisfy (1.2). Clearly, it suffices to consider (1.2) with $y = 0$, $x = 0$, $x' = 1$. Then

$$(1-r)U(0|1) + rU(1|1) = rU(0|0) + rU(1|0)$$

which is satisfied by exactly those channels U for which $U(1|1) = 1$. Thus $\Lambda_0(P) = p$, $\Lambda_0 = \max_p \Lambda_0(P) = 1$, and from (1.7) and (3.4) we obtain

$$C(\Lambda) = \max_{p:\, p \ge \Lambda} I(p,\Lambda) \tag{3.5}$$

for every $\Lambda < 1$. On the other hand, the random code capacity $C_r(\Lambda)$ equals $\max_p I(p,\Lambda)$, with no constraint on p (cf. (1.10)).

Equation (3.3) offers a geometric interpretation which can be used to obtain $I(p,\Lambda)$ as shown in Fig. 1. Let A, B, and D be points on the $h(t)$ curve with $t_1 = r$, $t_2 = r + (1-\Lambda)(1-2r)$, and $t = r + (1-p)(1-\Lambda)(1-2r)$, respectively. Then $I(p,\Lambda)$ equals that part of the ordinate of D which lies above the secant AB. Further, with increasing p, the point D moves from right to left. Clearly, $I(p,\Lambda)$ attains its maximum if and only if the tangent to the $h(t)$ curve at D is parallel to the secant AB. Denoting the maximizing p by $p^* = p^*(\Lambda, r)$, we observe that $I(p,\Lambda)$ increases with p for $p \le p^*$ and decreases for $p > p^*$ (explicit formulas for p^* and $C_r(\Lambda) = I(p^*, \Lambda)$, though easy to obtain, are omitted owing to their tediousness). Thus

$$C(\Lambda) = C_r(\Lambda) \quad \text{if } \Lambda \le p^*(\Lambda, r)$$
$$C(\Lambda) = I(\Lambda,\Lambda) < C_r(\Lambda) \text{ if } \Lambda > p^*(\Lambda, r). \tag{3.6}$$

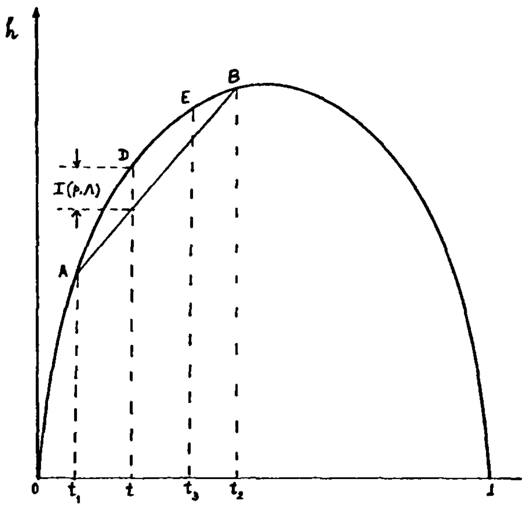

Fig. 1. $t_1 = r$. $t_2 = r + (1-\Lambda)(1-2r)$. $t_3 = r + (1-\Lambda)^2(1-2r)$. $t = r + (1-p)(1-\Lambda)(1-2r)$.

Geometrically (see Fig. 1), the first or second case will obtain accordingly to whether the tangent to the $h(t)$ curve at the point E, corresponding to $t_3 = r + (1-\Lambda)^2(1-2r)$, has a smaller (possibly equal) or a larger slope, respectively, than the secant AB. In other words, $\Lambda \le p^*(\Lambda, r)$ if and only if

$$\log \frac{1 - r - (1-\Lambda)^2(1-2r)}{r + (1-\Lambda)^2(1-2r)} \le \frac{h(r + (1-\Lambda)(1-2r)) - h(r)}{(1-\Lambda)(1-2r)}. \tag{3.7}$$

We now show that (3.7) necessarily holds if $\Lambda \le 1/2$. To this end, observe in Fig. 1 that always

$$t_1 < t_2 \le 1 - t_1 \tag{3.8}$$

and $t_3 \ge (t_1 + t_2)/2$ if $\Lambda \le 1/2$. Hence (by concavity) it suffices to show that (3.8) implies

$$h'\left(\frac{t_1 + t_2}{2}\right) \le \frac{h(t_2) - h(t_1)}{t_2 - t_1}$$

or, substituting $t = (t_1 + t_2)/2$, $u = (t_2 - t_1)/2$, that

$$h(t+u) - h(t-u) - 2uh'(t) \ge 0, \qquad \text{if } 0 \le u \le t \le \tfrac{1}{2}.$$

The last inequality holds with equality for $u = 0$, and differentiation shows that the left side is an increasing function of u in the interval $0 \le u \le t$. This establishes our claim, i.e., for $\Lambda \le 1/2$ we always have $C(\Lambda) = C_r(\Lambda)$.

Next, as an application of Theorems 2 and 5 we prove that the capacity of the OR channel under state constraint Λ can be attained by the MD decoding rule. First we show that the hypothesis of Theorem 2 is satisfied by any $P = (1-p, p)$ with $p > \Lambda$, i.e., we have for every $Q = (1-q, q)$ with $q \le \Lambda$ and random variables X, S, Y with $P_{XSY}(x,s,y) = P(x)Q(s)Q(y|x,s)$, that

$$I(X \wedge Y) > I(S \wedge Y). \tag{3.9}$$

Now, $I(X \wedge Y) = I(p,q)$ (cf. (3.3)) and by symmetry,

$I(S \wedge Y) = I(q, p)$. Then by (3.3),

$$\begin{aligned} I(X \wedge Y) - I(S \wedge Y) &= -(1-p)h(r+(1-q)(1-2r)) - ph(r) \\ &\quad + (1-q)h(r+(1-p)(1-2r)) + qh(r) \\ &= (1-p)(1-q)(1-2r)\left[\frac{h(r+(1-p)(1-2r)) - h(r)}{(1-p)(1-2r)} \right. \\ &\quad \left. - \frac{h(r+(1-q)(1-2r)) - h(r)}{(1-q)(1-2r)}\right]. \end{aligned}$$

In the square brackets the first of the two difference quotients is larger since $h(t)$ is concave and $(1-p)(1-2r) < (1-q)(1-2r)$ by the assumption $p > \Lambda \geq q$. This establishes (3.9). To prove the goodness of the MD decoding rule for all $P = (1-p, p)$ with $p > \Lambda$, it suffices in view of Theorem 5 to show that $W_Q(1|0) + W_Q(0|1) < 1$ for every $Q = (1-q, q)$ with $q \leq \Lambda$. This, however, is obvious from (3.2) even without the constraint $q \leq \Lambda$. The following theorem summarizes our main result for the OR channel.

Theorem 6: The capacity of the OR channel with additive binary $(1-r, r)$ noise $(0 \leq r < 1/2)$ and state constraint Λ $(0 < \Lambda < 1)$ is given by

$$C(\Lambda) = \begin{cases} \max_p I(p, \Lambda) = C_r(\Lambda), & \text{if (3.7) holds} \\ I(\Lambda, \Lambda) < C_r, & \text{otherwise} \end{cases} \tag{3.10}$$

with $I(p, \Lambda)$ defined as in (3.3). In particular, for $\Lambda \leq 1/2$, always $C(\Lambda) = C_r(\Lambda)$. Finally, the capacity (3.10) can be attained by the minimum distance decoding rule.

For the noiseless OR channel (i.e., for $r = 0$) we will show in Appendix III that the random code capacity equals

$$C_r(\Lambda) = \log\left[1 + (1-\Lambda)\Lambda^{\Lambda/1-\Lambda}\right] \tag{3.11}$$

and that (3.7) is satisfied if and only if $\Lambda \leq \Lambda^* = 0.6086$. Hence for a noiseless OR channel the capacity formula (3.10) becomes

$$C(\Lambda) = \begin{cases} \log\left[1 + (1-\Lambda)\Lambda^{\Lambda/1-\Lambda}\right] = C_r(\Lambda), \\ \qquad\qquad \text{if } \Lambda \leq \Lambda^* \\ h((1-\Lambda)^2) - (1-\Lambda)h(\Lambda) < C_r(\Lambda), \\ \qquad\qquad \text{if } \Lambda > \Lambda^*. \end{cases} \tag{3.12}$$

Numerical computations indicate that also for $r > 0$ there exists a threshold $\Lambda^*(r)$ such that $C(\Lambda)$ equals the random code capacity if and only if $\Lambda \leq \Lambda^*(r)$; moreover, $1/2 \leq \Lambda^*(r) \leq \Lambda^* = 0.6086$ and $\Lambda^*(r) \to 1/2$ if $r \to 1/2$. While we do not have a proof for these results, we will prove in Appendix III that for any given $r < 1/2$, $C(\Lambda)$ is strictly less than the random code capacity if Λ is sufficiently close to 1.

We know that the capacity of any OR channel can be attained by the MD decoding rule. In the noiseless case the simplest decoding rule suffices: decode message i if the received sequence $\boldsymbol{y}$ could have arisen due to codeword $\boldsymbol{x}_i$ (i.e., if $\boldsymbol{x}_i$ has a zero in every position where $\boldsymbol{y}$ has a zero); if there are more such codewords than one, declare an error. This decoding rule clearly improves upon "minimum distance" as the latter could result in an impossible codeword; otherwise, the two rules are equivalent.

Our result for the noiseless OR channel affords the following combinatorial interpretation. Let $N(n, \Lambda, \epsilon)$ denote the maximum number N of binary n sequences such that if $\lfloor n\Lambda \rfloor$ positions are covered with the covered bits being always read as 1s, then regardless of which positions are covered it is still possible to identify at least $N(1-\epsilon)$ of these sequences (the nonidentifiable sequences may depend on the positions covered). Then

$$\lim_{n \to \infty} \frac{1}{n} \log N(n, \Lambda, \epsilon) = C(\Lambda), \qquad \text{for all } 0 < \epsilon < 1$$

with $C(\Lambda)$ given by (3.12). The same problem appears much more difficult when $\epsilon = 0$ (compare with the discussion of [9, Example 1]. If the positions were instead covered at random, each with probability Λ, and identifiability were required with high probability, this becomes a standard coding problem for a discrete memoryless channel known as the Z-channel. The capacity of the latter channel equals $C_r(\Lambda)$ in (3.11). Note that for $\Lambda > \Lambda^*$, the deterministic and random versions of this combinatorial problem have different answers, which is at variance with [9, Example 1].

IV. Group Adder AVC's

In this section we consider AVC's whose inputs, states and outputs are elements of a commutative group $\mathscr{G}$, the output being determined by the group addition of input, state, and possibly noise. Formally, such an AVC is defined by

$$W(y|x,s) = R(y - x - s) \tag{4.1}$$

where R is a given distribution on $\mathscr{G}$, called the noise distribution. These AVC's, called group adder AVC's, constitute a special class of additive AVC's (cf. (2.13)) treated in the next section. The group $\mathscr{G}$ may be finite or infinite (e.g., $\mathscr{G} = \mathbb{R}^d$, with ordinary vector addition), but $\mathscr{X}$, $\mathscr{S}$, and $\mathscr{Y}$ are restricted to be finite subsets of $\mathscr{G}$, and R is assumed to have a finite support.

A noiseless group adder AVC is obtained when R is chosen as the point mass at the 0 element of $\mathscr{G}$. Two simple examples of such an AVC are the binary adder and arithmetic adder AVC's in [9, Examples 1 and 2].

We assume that $0 \in \mathscr{X}$, $0 \in \mathscr{S}$, and consider state constraints of the form (1.4) with an arbitrary function l satisfying $l(s) \geq l(0) = 0$. Further, we assume that the noise distribution R is *regular* in the sense of (2.16), i.e.,

$$Q_1 * R = Q_2 * R \text{ implies } Q_1 = Q_2. \tag{4.2}$$

Clearly, the AVC defined by (4.1) is symmetric if $\mathscr{X} = \mathscr{S}$; we will show that it is nonsymmetrizable if $|\mathscr{S}| < |\mathscr{X}|$. Further, we will provide a formula for the key functional

$\Lambda_0(P)$ (cf. (1.5)), in general. To this end, we first identify the set $\mathcal{U}$ of channels satisfying (1.2).

Substituting (4.1) into (1.2), we obtain the equations

$$\sum_{s\in\mathcal{G}} R(y-x-s)U(s|x') = \sum_{s\in\mathcal{G}} R(y-x'-s)U(s|x) \tag{4.3}$$

with the understanding that if $\mathcal{S}$ is a proper subset of $\mathcal{G}$, then $U(s|x)=0$ for $s\notin\mathcal{S}$. Setting $x'=0$, and substituting on the left side $x+s=t$, (4.3) yields

$$\sum_{t\in\mathcal{G}} R(y-t)U(t-x|0) = \sum_{s\in\mathcal{G}} R(y-s)U(s|x). \tag{4.4}$$

Denoting the distribution $U(\cdot|0)$ by U_0, (4.4) states that the convolutions of R with the translate of U_0 by x, and with $U(\cdot|x)$, respectively, are equal. Then by the regularity assumption (4.2),

$$U(s|x) = U_0(s-x) \tag{4.5}$$

for every $x\in\mathcal{X}$, $s\in\mathcal{G}$.

If $\mathcal{S}=\mathcal{G}$, then any channel U: $\mathcal{X}\to\mathcal{S}$ of the form (4.5), for any distribution U_0 on $\mathcal{G}$, obviously belongs to $\mathcal{U}$ thereby determining $\mathcal{U}$ for this case. If $\mathcal{S}$ is a proper subset of $\mathcal{G}$, the distribution U_0 in (4.5) is constrained by the condition that $U(s|x)=0$ if $s\notin S$, $x\in\mathcal{X}$. Thus $U_0(t)$ must be 0 for every $t\in\mathcal{G}$ that can be represented as $t=s-x$ with $s\notin\mathcal{S}$, $x\in\mathcal{X}$. In other words, a necessary condition for $U\in\mathcal{U}$ is that it be of the form (4.5) with a distribution on $\mathcal{G}$ such that

$$U_0(t)=0 \text{ whenever } x+t\notin S \text{ for some } x\in\mathcal{X}. \tag{4.6}$$

It is easy to see that this condition is also sufficient. Notice that (4.6) implies that U_0 is supported on $\mathcal{S}$, because for $t\notin\mathcal{S}$ the choice $x=0$ yields $U(t)=0$.

If $|\mathcal{S}|<|\mathcal{X}|$, then for every $t\in\mathcal{G}$ there exists $x\in\mathcal{X}$ with $x+t\notin\mathcal{S}$, and hence no distribution can satisfy (4.6). Thus in this case $\mathcal{U}=\phi$ as claimed.

Having determined $\mathcal{U}$ for a group adder AVC in general, we obtain from (1.5) that

$$\begin{aligned}\Lambda_0(P) &= \min_{U\in\mathcal{U}} \sum_{x,s} P(x)U(s|x)l(s)\\ &= \min_{U_0} \sum_{x,s} P(x)U_0(s-x)l(s) = \min_{U_0} l(P*U_0)\end{aligned} \tag{4.7}$$

where the minimum is taken over all distributions U_0 on $\mathcal{S}$ satisfying (4.6), and where $l(\cdot)$ for a distribution is defined by (1.9), i.e., $l(Q)=\Sigma_s Q(s)l(s)$.

The capacity of a general AVC is determined in terms of the mutual information functional $I(P,W_Q)$ (cf. (1.3), (1.7), (1.8)), where $I(P,W_Q)$ denotes the mutual information $I(X\wedge Y)$ for X and Y with $P_{XY}(x,y)=P(x)W_Q(y|x)$. For a group adder AVC (4.1), the last condition means that

$$\begin{aligned}&Y=X+S+Z, \qquad \text{with } X,S,Z \text{ independent},\\ &P_X=P,\ P_S=Q,\ P_Z=R.\end{aligned} \tag{4.8}$$

From (4.8), we can write

$$I(P,W_Q)=H(Y)-H(S+Z). \tag{4.9}$$

We now determine the AVC capacity under state constraint Λ when $\mathcal{X}=\mathcal{G}$. Then if $\mathcal{S}$ is a proper subset of $\mathcal{G}$, the AVC is nonsymmetrizable and hence $\Lambda_0(P)$ is identically $+\infty$. For the case $\mathcal{S}=\mathcal{G}$, we need the following lemma.

Lemma 3: In the case $\mathcal{X}=\mathcal{S}=\mathcal{G}$, $\max_P \Lambda_0(P)$ is attained by the uniform distribution $P^*=\{P(x)=1/|\mathcal{G}|\}_{x\in\mathcal{G}}$, and

$$\Lambda_0=\max_P \Lambda_0(P)=\Lambda_0(P^*)=\frac{1}{|\mathcal{G}|}\sum_{s\in\mathcal{G}} l(s).$$

Proof: In this case, $\mathcal{U}$ consists of all channels of the form (4.5) with no restriction on U_0, and from (4.7)

$$\max_P \Lambda_0(P)=\max_P \min_{U_0} l(P*U_0).$$

Since the convolution of the uniform distribution with any other distribution on $\mathcal{G}$ is again the uniform distribution, (P^*,P^*) is a saddle point of $l(P*U_0)$.

Theorem 7: The capacity of a group adder AVC (4.1) with $\mathcal{X}=\mathcal{G}$ under state constraint $\Lambda<\Lambda_0$ (given by Lemma 3 if $\mathcal{S}=\mathcal{G}$, and $\Lambda_0=\infty$ if $\mathcal{S}$ is a proper subset of $\mathcal{G}$) is

$$C(\Lambda)=\log|\mathcal{G}|-\max_{Q:\ l(Q)\le\Lambda} H(Q*R) \tag{4.10}$$

while for $\Lambda\ge\Lambda_0$ the capacity of 0. Further, the capacity can be attained by the MMI decoding rule.

Corollary: The capacity of this AVC without state constraint is 0 if $\mathcal{S}=\mathcal{G}$ and is positive and equal to

$$C=\log|\mathcal{G}|-\max_Q H(Q*R)$$

if $\mathcal{S}$ is a proper subset of $\mathcal{G}$.

Proof: By (1.7) and (1.8), for $\Lambda<\Lambda_0$ we have

$$C(\Lambda)=\max_{P:\ \Lambda_0(P)\ge\Lambda}\ \min_{Q:\ l(Q)\le\Lambda} I(P,W_Q)$$

where $I(P,W_Q)$ is now given by (4.8) and (4.9). The max and min can be interchanged by the standard min–max theorem argument. For a fixed Q the maximum of $I(P,W_Q)$ is achieved for a choice of P that maximizes $H(Y)$, and $\max H(Y)=\log|\mathcal{G}|$ is attained when $P=P^*$, the uniform distribution on $\mathcal{G}$. Notice that P^* satisfies the constraint $\Lambda_0(P^*)\ge\Lambda$ by Lemma 3. This completes the proof of the capacity formula (4.10). For $\Lambda\ge\Lambda_0$, the capacity is obviously 0 because $Q=P^*$ satisfies the constraint $l(Q)\le\Lambda$, and yields $I(P,W_Q)=0$ for all P.

To prove the last assertion, on account of Theorem 2 it suffices to check that for random variables as in (4.8), with $P=P^*$ and $l(Q)\le\Lambda$

$$I(X\wedge Y)\ge I(S\wedge Y).$$

Since (4.8) implies $I(S\wedge Y)=H(Y)-H(X+Z)$, for $P=P^*$ we have $I(S\wedge Y)=0$, while $I(X\wedge Y)>C(\Lambda)>0$. This completes the proof.

Remarks: 1) Theorem 7 shows, in particular, that in this case the capacity is equal to the random code capacity.

Notice the remarkable fact that in Theorem 7 no assumption was needed on the function l appearing in the state constraint.

2) Setting $\mathcal{G} = \{0,1\}$ with mod-2 addition, and $l(s)=s$, Theorem 7 covers the noiseless binary adder AVC of [9] as well as its noisy version with arbitrary "noise distribution" $R=(1-r,r)$, $r \neq 1/2$. For this channel Theorem 7 gives the capacity $C(\Lambda)=1-h(\Lambda * r)$ if $\Lambda<1/2$ (and 0 if $\Lambda \geq 1/2$), which for $r=0$ reduces to $1-h(\Lambda)$ as obtained in [9]. Notice that by applying Theorem 5, it follows that the capacity of this binary AVC can be attained by the MD decoding rule provided that $r<1/2$.

For the case where $\mathcal{X}$ is a proper subset of $\mathcal{G}$, no result as explicit as Theorem 7 is expected. Nevertheless, a general result will be stated as Theorem 8. It will involve capacity under input constraint Γ (in addition to state constraint Λ), with the understanding that the input and state constraints (1.4) and (1.11) are defined in terms of the *same* function $l(\cdot)$,

Theorem 8: For a group adder AVC with state constraint Λ, the typicality decoding rule is good for every strictly positive input distribution P satisfying $\Lambda_0(P)>\Lambda$. For those P's as before which are regular, the independence decoding rule is also good. Moreover, if either 1) $\mathcal{X}=\mathcal{S}$, and this set is not closed under addition in $\mathcal{G}$, or 2) $\mathcal{G}=\mathbb{R}^d$, $\mathcal{X} \subset \mathcal{S} \subset (\mathbb{R}^+)^d$, and $l(\cdot)$ is a nondecreasing function in the sense that $f(t_1) \leq f(t_2)$ if $t_1 \leq t_2$ componentwise, then

$$\Lambda_0(P)=l(P), \qquad \Lambda_0=\max_{x\in\mathcal{X}} l(x) \tag{4.11}$$

and

$$C(\Gamma,\Lambda)=\max_{P:\ \Lambda\leq l(P)\leq\Gamma} I(P,\Lambda)>0 \tag{4.12}$$

if $\Gamma>\Lambda$, $\Lambda<\Lambda_0$, while $C(\Gamma,\Lambda)=0$, otherwise.

Remark: $I(P,\Lambda)$ is concave in P, being the minimum of concave functions $I(P,W_Q)$. Hence the right side of (4.12) is either equal to $C_r(\Lambda)=\max_P I(P,\Lambda)$, or else the maximum is attained at the boundary. More precisely, if $I(P^*,\Lambda)=C_r(\Lambda)$ implies $l(P^*)<\Lambda$, then the maximum is attained with $l(P)=\Lambda$, whereas if $l(P^*)>\Gamma$, it is attained with $l(P)=\Gamma$. All these cases are actually possible as already shown by the simple examples of [9].

Proof: To prove the first assertions, on account of Theorems 3 and 4 it suffices to show that if $\Lambda_0(P)>\Lambda$, then (2.9) and (2.10) cannot simultaneously hold. Suppose, therefore, that (2.9) holds for some Q and U. With (4.1), this means that

$$\sum_{x,s} P(x)R(y-x-s)U(s|x')=\sum_{s} R(y-x'-s)Q(s) \tag{4.13}$$

where as in (4.3) we extend the summation over s to all of $\mathcal{G}$ by setting $U(s|x)=Q(s)=0$ for $s\notin\mathcal{S}$, while x ranges over $\mathcal{X}$. Substituting $x'=0$, we obtain

$$R*P*U_0=R*Q \quad (\text{where } U_0(\cdot)=U(\cdot|0))$$

implying that $P*U_0=Q$, i.e.,

$$\sum_{x} P(x)U_0(s-x)=Q(s). \tag{4.14}$$

Further, for any $t\in\mathcal{G}$ such that $t+x'\notin\mathcal{S}$ for some $x'\in\mathcal{X}$, it follows from (4.14) with the substitution $s=t+x'$ that $U_0(t)=0$ (using the assumed strict positivity of P). Hence (4.14) yields that $Q=P*U_0$ for some U_0 satisfying (4.6). On account of (4.7), we then get $\Lambda_0(P)\leq l(Q)$. By the assumption $\Lambda_0(P)>\Lambda$, this means that (2.10) does not hold. This completes the proof of the first assertions of the theorem.

Further, in case 1), it is clear from (4.6) that the unit mass at 0 is the only admissible U_0; thus, trivially $\Lambda_0(P)=l(P)$. In case 2), for any admissible U_0,

$$\begin{aligned} l(P*U_0) &= \sum_{x,s} P(x)U_0(s-x)l(s) \\ &\geq \sum_{x,s} P(x)U_0(s-x)l(x)=l(P) \end{aligned}$$

where the inequality follows from the monotonicity of l since $U_0(s-x)>0$ implies $s\geq x$. Hence by (4.7), $\Lambda_0(P)=l(P)$ also in this case, proving (4.11). The last assertions follow from (4.11) by (1.12)—applied now with $g(\cdot)=l(\cdot)$—together with the passage following (1.12).

We conclude this section by observing that the key to our treatment of group adder AVCs was the explicit description (viz (4.5), (4.6)) of the set $\mathcal{U}$ of channels satisfying (1.2). Under suitable conditions on $l(\cdot)$ an alternative approach is possible, which applies also to general additive AVC's as will be seen in the next section.

V. Additive AVC's

In this section we will study additive AVC's W whose input and output alphabets are (finite) subsets of $\mathbb{R}^d$. Thus according to (2.13),

$$W(y|x,s)=V(y-x|s) \tag{5.1}$$

where the set of states $\mathcal{S}$ is an arbitrary finite set, and V: $\mathcal{S}\to\mathcal{Z}$ is some given channel from $\mathcal{S}$ into a finite set $\mathcal{Z}\subset\mathbb{R}^d$. For $z\notin\mathcal{Z}$, we set $V(z|s)=0$; with this convention, (5.1) makes sense also when $y-x\notin\mathcal{Z}$.

The model (5.1) includes the extension of the group adder AVC (with $\mathcal{G}=\mathbb{R}^d$) to the case where the noise distribution may also vary arbitrarily. To see this, set $\mathcal{S}=\mathcal{S}_1\times\mathcal{S}_2$ where $\mathcal{S}_1\subset\mathbb{R}^d$ and to each $s_2\in\mathcal{S}_2$ there corresponds a "noise distribution" R_{s_2} on $\mathbb{R}^d$ (with finite support). Then setting $V(z|s_1,s_2)=R_{s_2}(z-s_1)$, (5.1) gives the following generalization of (4.1):

$$W(y|x,s_1,s_2)=R_{s_2}(y-x-s_1). \tag{5.2}$$

For AVC's of this kind, it is natural to let the "cost" depend on s_1 alone and also to postulate that all "noise distributions" R_{s_2} have zero expectation vector. Then the expectation vector of $V(\cdot|s_1,s_2)$ defined by $V(z|s_1,s_2)=R_{s_2}(z-s_1)$ equals s_1, and the cost $l(s)$ depends on $s=(s_1,s_2)$ through this expectation vector only.

For general additive AVC's (cf. (5.1)), we will denote the expectation vector of $V(\cdot|s)$ by E_s, i.e.,

$$E_s = \sum_{z \in \mathscr{Z}} V(z|s)z. \tag{5.3}$$

Motivated by the foregoing consideration, we restrict ourselves to state constraints Λ with $l(s)$ depending on s through E_s. Thus with some abuse of notation,

$$l(s) = l(E_s)$$

where the latter l is a nonnegative function on $\mathbb{R}^d$ with $l(\mathbf{0}) = \mathbf{0}$. Input constraints Γ will be understood in terms of the same function $l(\cdot)$. Inequalities for elements of $\mathbb{R}^d$ will be understood as holding componentwise.

Theorem 9: For an additive AVC with input and state constraints as just given, with $\Lambda < \max_{x \in \mathscr{X}} l(x)$, suppose either

1) $\mathscr{X} \subset (\mathbb{R}^+)^d$, $\mathbf{0} \in \mathscr{X}$, $E_s \geq \mathbf{0}$ for all $s \in \mathscr{S}$, and $l(\cdot)$ is convex and $l(t_1) \leq l(t_2)$ for $t_1 \leq t_2$; or
2) $\mathscr{X}$ and $\mathscr{Z}$ are symmetric around $\mathbf{0}$, for each $s \in \mathscr{S}$ there is an $s' \in \mathscr{S}$ such that $V(z|s') = V(-z|s)$ for all $z \in \mathscr{Z}$, and $l(\cdot)$ is a convex even function.

Under condition 1 (respectively, 2), the independence decoding rule is good for every (respectively every symmetric) distribution P on $\mathscr{X}$ with $\min_x P(x) > 0$ and $l(P) > \Lambda$. In particular, in both cases

$$C(\Gamma, \Lambda) \geq \max_{P:\ \Lambda \leq l(P) \leq \Gamma} I(P, \Lambda), \qquad \text{if } \Gamma > \Lambda. \tag{5.4}$$

If, in addition, there exists a mapping f: $\mathscr{X} \to \mathscr{S}$ such that for all $x \in \mathscr{X}$, $z \in \mathbb{R}^d$,

$$V(z|f(x)) = R(z - x) \tag{5.5}$$

where R is a distribution on $\mathbb{R}^d$ with zero expectation vector, then (5.4) holds with equality and moreover $C(\Gamma, \Lambda) > 0$ if and only if $\Gamma > \Lambda$.

Remark: Hypothesis (5.5) implies that the AVC W is deterministically symmetrizable. In fact, by (5.5) with $z = y - x'$, (5.1) gives that

$$W(y|x', f(x)) = V(y - x'|f(x)) = R(y - x - x');$$

by symmetry, $W(y|x, f(x'))$ equals the same. Observe that hypothesis (5.5) always holds for the AVC's described in the paragraph containing (5.2).

Proof: By Theorem 4 the goodness of the independence decoding rule will be established if we demonstrate that P satisfies DS (Λ), i.e., that (2.9) and (2.10) cannot be simultaneously satisfied.

For W as in (5.1), (2.9) becomes

$$\sum_{x,s} P(x)V(y - x|s)U(s|x') = \sum_s Q(s)V(y - x'|s). \tag{5.6}$$

We will show that (5.6) implies

$$\sum_{x,s} P(x)U(s|x)l(E_s) > \Lambda, \tag{5.7}$$

i.e., (recalling that now $l(s) = l(E_s)$) that the second inequality in (2.10) can never hold if (2.9) does.

To this end, multiply both sides of (5.6) by y and sum over y. Since with the substitution $y = x + z$,

$$\sum_y V(y - x|s)y = x + \sum_x V(z|s)z = x + E_s$$

it follows that

$$\sum_x P(x)x + \sum_s U(s|x')E_s = x' + \sum_s Q(s)E_s. \tag{5.8}$$

This, together with the convexity of $l(\cdot)$, yields

$$\begin{aligned}\sum_{x',s} & P(x')U(s|x')l(E_s) \\ &\geq \sum_{x'} P(x')l\Big(\sum_s U(s|x')E_s\Big) \\ &= \sum_{x'} P(x')l\Big(x' + \sum_s Q(s)E_s - \sum_x P(x)x\Big). \end{aligned} \tag{5.9}$$

In case 1), substituting $x' = \mathbf{0}$ in (5.8), we see that $\sum_x P(x)x \leq \sum Q(s)E_s$. Therefore, by the assumption on $l(\cdot)$, the last sum in (5.9) is bounded below by $\sum_{x'} P(x')l(x') = l(P)$. By the hypothesis $l(P) > \Lambda$, this establishes (5.7).

In case 2), since l is now a convex even function, we have

$$\begin{aligned} l(x + c) + l(-x + c) &= l(x + c) + l(x - c) \\ &\geq 2l(x) = l(x) + l(-x) \end{aligned}$$

for every $c \in \mathbb{R}^d$. Hence for a symmetric P

$$\begin{aligned}\sum_x P(x)l(x + c) &= \frac{1}{2}\sum_x P(x)(l(x + c) + l(-x + c)) \\ &\geq \frac{1}{2}\sum_x P(x)(l(x) + l(-x)) \\ &= \sum_x P(x)l(x).\end{aligned}$$

This means that in this case too, the last sum in (5.9) is bounded below by $l(P) > \Lambda$, completing the proof of the first assertion of Theorem 9.

Next, if P is any input distribution with $l(P) < \Gamma$ for which a good decoding rule exists (given the state constraint Λ), then by definition $C(\Gamma, \Lambda) > I(P, \Lambda)$. In particular, in case 1) the first assertion (just proved) of the theorem immediately implies inequality (5.4). In case 2), the same would follow if we showed that the maximum in (5.4) is attained for a symmetric P. This, however, is a consequence of the fact that hypothesis 2) implies $I(P, \Lambda) = I(P', \Lambda)$ whenever $P'(x) = P(-x)$ for every $x \in \mathscr{X}$, and of the concavity of $I(P, \Lambda)$ as a function of P.

Finally, under hypothesis (5.5), the deterministic channel U: $\mathscr{X} \to \mathscr{S}$ defined by f belongs to the set U of channels satisfying (1.2), as observed in the earlier remark. Since (5.5) and (5.3) imply $E_{f(x)} = x$ (because R has $\mathbf{0}$ expectation), it follows that

$$\begin{aligned}\Lambda_0(P) &= \min_{U \in \mathscr{U}} \sum_{x,s} P(x)U(s|x)l(E_s) \\ &\leq \sum_x P(x)l(E_{f(x)}) = l(P).\end{aligned}$$

This result leads, by (1.9) with $g(\cdot)=l(\cdot)$, to the conclusion that (5.4) must hold with equality, and $C(\Gamma,\Lambda)>0$ if $\Gamma>\Lambda$. By the argument following (1.12), it also follows that $C(\Gamma,\Lambda)=0$ if $\Gamma\le\Lambda$. This completes the proof of Theorem 9.

VI. Randomized State Selection

In this section, we discuss the capacity of an AVC when the state selector is permitted to randomize. Recall that for a code with codewords $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N$ and decoder ϕ, the average probability of error for a given state sequence $\boldsymbol{s}\in\mathcal{S}^n$ is

$$\bar{e}(\boldsymbol{s})=\frac{1}{N}\sum_{i=1}^{N}\sum_{\boldsymbol{y}:\,\phi(\boldsymbol{y})\ne i}W^n(\boldsymbol{y}|\boldsymbol{x}_i,\boldsymbol{s}). \tag{6.1}$$

If randomized state selection is permitted subject to certain constraints, a positive number R is called an ϵ-achievable rate if for every $\delta>0$ and sufficiently large n there exist codes with rate larger than $R-\delta$ and with $E\bar{e}(\boldsymbol{S})<\epsilon$ uniformly for all admissible random state sequences $\boldsymbol{S}=(S_1,\cdots,S_n)$. The corresponding capacity is the largest R which is ϵ-achievable for every $\epsilon>0$.

The capacity of an AVC without state constraints is not affected by randomized state selection because any code that satisfies $\bar{e}(\boldsymbol{s})<\epsilon$ for every $\boldsymbol{s}\in\mathcal{S}^n$ also satisfies $E\bar{e}(\boldsymbol{S})<\epsilon$ for every random state sequence $\boldsymbol{S}$. Similarly, the capacity under state constrained Λ remains unchanged by randomized state selection subject to the constraint

$$l(\boldsymbol{S})\le\Lambda \text{ almost surely.} \tag{6.2}$$

On the other hand, if (6.2) is replaced by the "average constraint"

$$El(\boldsymbol{S})\le\Lambda \tag{6.3}$$

then the corresponding capacity will be equal to that of the unconstrained AVC (cf. Csiszár–Narayan [8]; that random codes were considered there does not make any difference in the present context). Also, as in [8], under the constraint (6.3) the ϵ-capacity (the largest ϵ-achievable rate) is typically larger, for any fixed $\epsilon>0$, than its limit as $\epsilon\to0$ which is (by definition) the capacity.

To obtain more attractive results, randomized state selections subject to (6.3) will be addressed with the additional restriction that the states at different instants be selected independently, i.e., the components of $\boldsymbol{S}=(S_1,\cdots,S_n)$ be independent random variables. Three different models are considered.

1) $S_1,\cdots,S_n$ are required to have the same distribution; then (6.3) reduces to the constraint

$$l(Q)\le\Lambda \tag{6.4}$$

on the common distribution Q of $S_1,\cdots,S_n$.

2) $S_1,\cdots,S_n$ need not have the same distribution, but each should satisfy

$$El(S_i)\le\Lambda. \tag{6.5}$$

3) No restrictions, other than independence and (6.3), are imposed on $S_1,\cdots,S_n$.

In each case we will also suppose that an input constraint Γ is given in terms of some "cost function" $g(\cdot)$ on $\mathcal{X}$ (cf. (1.11); here, unlike in the previous section, $g(\cdot)$ is arbitrary and may be unrelated to $l(\cdot)$).

Theorem 10: The capacity under input constraint Γ of an AVC with random state selection as in 1)–3), abbreviated $C^{(i)}(\Gamma,\Lambda)$, $i=1,2,3$, is given by

$$C^{(1)}(\Gamma,\Lambda)=C_r(\Gamma,\Lambda)$$

$$C^{(2)}(\Gamma,\Lambda)=\begin{cases}C_r(\Gamma,\Lambda)>0, & \text{if } \Lambda<\Lambda_0\\ 0, & \text{if } \Lambda\ge\Lambda_0\end{cases}$$

$$C^{(3)}(\Gamma,\Lambda)=\begin{cases}C(\Gamma,\Lambda)>0, & \text{if } \Lambda<\max\limits_{P:\,g(P)\le\Gamma}\Lambda_0(P)\\ 0, & \text{if } \Lambda\ge\max\limits_{P:\,g(P)\le\Gamma}\Lambda_0(P)\end{cases}$$

with $C_r(\Gamma,\Lambda)$ and $C(\Gamma,\Lambda)$ as in (1.12), (1.13).

Proof: If $S_1,\cdots,S_n$ are independent state random variables with distributions $Q_1,\cdots,Q_n$, then with the notation

$$W^n_{\mathbf{Q}}(\mathbf{y}|\mathbf{x})=\prod_{i=1}^{n}W_{Q_i}(y_i|x_i) \tag{6.6}$$

(cf. (1.1)), we obtain from (6.1) that

$$\begin{aligned}E\bar{e}(\boldsymbol{S})&=\sum_{\boldsymbol{s}\in\mathcal{S}^n}Q_1(s_1)\cdots Q_n(s_n)\bar{e}(\boldsymbol{s})\\ &=\frac{1}{N}\sum_{i=1}^{N}\sum_{\boldsymbol{y}:\,\phi(\boldsymbol{y})\ne i}W^n_{\boldsymbol{Q}}(\boldsymbol{y}|\boldsymbol{x}_i)=\bar{e}(\boldsymbol{Q}).\end{aligned} \tag{6.7}$$

In other words, for any given code, $E\bar{e}(\boldsymbol{S})$ equals the average probability of error $\bar{e}(\boldsymbol{Q})$ of the same code on the AVC defined by (6.6) (whose states are the distributions on $\mathcal{S}$) for state sequence $\boldsymbol{Q}=(Q_1,\cdots,Q_n)$. The latter AVC is, in effect, the convex closure of the original one (cf., e.g., [7, p. 205]), but as different distributions Q may give rise to the same channel W_Q, the representation of the component channels of the convex closure as W_Q need not be unique.

In case 1), the sequence $\boldsymbol{Q}=(Q_1,\cdots,Q_n)$ consists of identical components satisfying (6.4). Hence using (6.7), the capacity $C^{(1)}(\Gamma,\Lambda)$ will be the same as the capacity under input constraint Γ of the compound channel defined by the family of channels W_Q with $l(Q)\le\Lambda$.

Compound channels are mathematically much simpler than AVC's. Indeed, by a trivial extension of the standard compound channel coding theorem (cf., e.g., [7, p. 173]) to the constrained input case, the capacity under input constraint Γ of the aforementioned compound channel is the maximum of

$$\min_{Q:\,l(Q)\le\Lambda}I(P,W_Q)=I(P,\Lambda) \tag{6.8}$$

over P satisfying $g(P)\le\Gamma$. This proves that $C^{(1)}(\Gamma,\Lambda)=C_r(\Gamma,\Lambda)$.

In case 2) we have a genuine AVC problem. By (6.5) and (6.7), $C^{(2)}(\Gamma,\Lambda)$ will be equal to the capacity under input constraint Γ of the AVC defined by $\{W_Q\colon l(Q)\le\Lambda\}$. The latter AVC must be considered without state constraints as the constraints (6.5) are fully taken into account by letting the state set be $\{Q\colon l(Q) \le \Lambda\}$. The inconvenience caused by this state set being infinite is easily overcome. (The hypothesis of a finite set of states could in general be dispensed with by an approximation argument such as in [7, p. 216]; this, however, is not needed now). Indeed, regarding the set $\{Q\colon l(Q) \le \Lambda\}$ as a simplex in the $|\mathcal{S}|$-dimensional space, let $\mathcal{S}^*$ denote the (finite) set of its vertices. Then the previous AVC is the convex closure of the AVC $\{W_Q\colon Q \in \mathcal{S}^*\}$ and hence has the same capacity under input constraint Γ as the latter. Since this capacity is equal to $C^{(2)}(\Gamma, \Lambda)$, it remains to prove that the last AVC (with finite set $\mathcal{S}^*$) is symmetrizable if and only if $\Lambda \ge \Lambda_0$, and that if nonsymmetrizable, its capacity under input constraint Γ equals $\max_{P\colon g(P) \le \Gamma} I(P, \Lambda) = C_r(\Gamma, \Lambda)$.

The last assertion follows from the fact that, in general, the capacity under input constraint Γ of a nonsymmetrizable AVC is the maximum over P (with $g(P) \le \Gamma$) of the minimum of the mutual information $I(P, V)$ for V in the convex closure of the given AVC (apply (1.12) with $\Lambda = l_{\max}$ to make the state constraint inoperative). As the convex closure of $\{W_Q\colon Q \in \mathcal{S}^*\}$ is $\{W_Q\colon l(Q) \le \Lambda\}$, the last minimum now equals (6.8). To verify the condition for symmetrizability, notice that any AVC (with finite state set) is symmetrizable if and only if its convex closure is deterministically symmetrizable. Since a mapping from $\mathcal{X}$ to $\{Q\colon l(Q) \le \Lambda\}$ is the same as a channel $U\colon \mathcal{X} \to \mathcal{S}$ with

$$\sum_s U(s|x) l(s) \le \Lambda, \qquad \text{for every } x \in \mathcal{X}, \tag{6.9}$$

the deterministic symmetrizability of the AVC $\{W_Q\colon l(Q) \le \Lambda\}$ means that

$$W_{U(\cdot|x')}(y|x) = W_{U(\cdot|x)}(y|x') \tag{6.10}$$

for some U satisfying (6.9). Recalling the definition (1.1) of W_Q, (6.10) means exactly that U satisfies (1.2), i.e., $U \in \mathcal{U}$. Finally, by (1.6), the existence of a $U \in \mathcal{U}$ satisfying (6.9) is equivalent to $\Lambda \ge \Lambda_0$. This completes the proof of Theorem 10 for case 2).

Case 3) could be dealt with similarly, but a direct approach is simpler. Since

$$C^{(3)}(\Gamma, \Lambda) \le C(\Gamma, \Lambda) \tag{6.11}$$

holds by definition for

$$\Lambda < \max_{P\colon g(P) \le \Gamma} \Lambda_0(P) \tag{6.12}$$

it suffices to establish the reverse inequality. Now, for an arbitrary sequence of independent random variables $\boldsymbol{S} = (S_1, \cdots, S_n)$ satisfying (6.3), Chebyshev's inequality implies that for any $\delta > 0$

$$\begin{aligned}\Pr\{l(\boldsymbol{S}) > \Lambda + \delta\} &\le \Pr\left\{\frac{1}{n}\sum_{i=1}^{n}[l(S_i) - El(S_i)] > \delta\right\}\\ &\le \frac{1}{n^2\delta^2}\sum_{i=1}^{n}\operatorname{var} l(S_i) \le l_{\max}^2/n\delta^2.\end{aligned}$$

Hence for any given code

$$E\bar{e}(\boldsymbol{S}) \le \max_{s\colon l(s) \le \Lambda + \delta} \bar{e}(s) + l_{\max}^2/n\delta^2.$$

It follows that $C^{(3)}(\Gamma, \Lambda) \ge C(\Gamma, \Lambda + \delta)$ for every $\delta > 0$, and since $C(\Gamma, \Lambda)$ is continuous in Λ subject to (6.12) (cf. (1.12)), this proves the reverse of inequality (6.11) under condition (6.12).

We still have to prove that if Λ does not satisfy (6.12), then $C^{(3)}(\Gamma, \Lambda) = 0$. Since by (1.5) and the minimax theorem

$$\max_{P\colon g(P) \le \Gamma} \Lambda_0(P) = \min_{U \in \mathcal{U}} \max_{P\colon g(P) \le \Gamma} \sum_{x,s} P(x) U(s|x) l(s),$$

if (6.12) does not hold then there exists $U \in \mathcal{U}$ such that

$$\sum_{x,s} P(x) U(s|x) l(s) \le \Lambda, \qquad \text{whenever } g(P) \le \Gamma. \tag{6.13}$$

Proceeding as in the proof of Lemma 1 of [9], consider any code with codewords $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ satisfying the input constraint (1.11), and let $\boldsymbol{S}_j = (S_{j1}, \cdots, S_{jn})$, $j = 1, \cdots, N$, be random state sequences with independent components whose distributions are defined by $\Pr\{S_{jk} = s\} = U(s|x_{jk})$, where $U \in \mathcal{U}$ satisfies (6.13). Then

$$El(\boldsymbol{S}_j) = \sum_{x,s} P_{\boldsymbol{x}_j}(x) U(s|x) l(s) \le \Lambda,$$

i.e., each $\boldsymbol{S}_j$ satisfies the constraint (6.3). On the other hand, by [9, eq. (3.29)],

$$E\bar{e}(\boldsymbol{S}_j) \ge \frac{N-1}{2N}, \qquad \text{for at least one } j.$$

This means that for no nontrivial code satisfying the initial input constraint (1.11) can $E\bar{e}(\boldsymbol{S})$ be uniformly small for all admissible state sequences $\boldsymbol{S}$; consequently,

$$C^{(3)}(\Gamma, \Lambda) = 0$$

In dealing with channels partially controlled by an adversary, McEliece [14] has considered a two-person zero-sum game between the "communicator" and the "jammer" with mutual information as the pay-off function. In our case this leads to the random code capacity

$$\begin{aligned}C_r(\Gamma, \Lambda) &= \max_{P\colon g(P) \le \Lambda} \min_{Q\colon l(Q) \le \Lambda} I(P, W_Q)\\ &= \min_{Q\colon l(Q) \le \Lambda} \max_{P\colon g(P) \le \Gamma} I(P, W_Q)\end{aligned} \tag{6.14}$$

as the value of the game. In justifying his approach, McEliece [14, pp. 134–135] remarks, first, that if the jammer were to use his optimal strategy, then no code with a rate higher than (6.14) could achieve a small average probability of error; second, the communicator could employ codes with rates arbitrarily close to (6.14) that ensure a small average probability of error regardless of the channel chosen by the jammer, *if this channel were memoryless*. The compound channel coding theorem is referred to as a reason for the last assertion which, therefore, must have been meant for the communication situation of our case 1). On the other hand, the rather vague formulation in

[14] could be interpreted as pertaining to cases 2) and 3) also. If true, this would indeed enhance the appeal of the mutual information game approach. By Theorem 10, when independent but not necessarily identically distributed jamming subject to (6.5) is permitted, the desired assertion is "almost true": it may occur that the jammer could prevent reliable transmission at any positive rate, but if not, the capacity remains equal to (6.14); the necessary and sufficient condition for the latter desirable case is $\Lambda < \Lambda_0$.

For independent jamming subject only to (6.3), the capacity may also be positive but strictly less than (6.14). Nevertheless, Theorem 10 permits us to identify those cases when indeed $C^{(3)}(\Gamma, \Lambda) = C_r(\Gamma, \Lambda)$; e.g., for additive AVC's such as in Section 5, this holds whenever $\Gamma > \Lambda$ and the unconstrained maximum of $I(P, \Lambda)$ is attained for some P with $l(P) \geq \Lambda$. Notice that $C^{(3)}(\Gamma, \Lambda)$, when positive, is always given as max–min of concave–convex functions over convex compact sets, viz.,

$$C(\Gamma, \Lambda) = \max_{\substack{P:\ \Lambda_0(P) \geq \Lambda \\ g(P) \leq \Gamma}} \ \min_{Q:\ l(Q) \leq \Lambda} I(P, W_Q)$$

(cf. (1.12) and (1.8)). Therefore, the max and min may be interchanged, and the Q^* attaining the minimum of

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

may be interpreted as an optimum strategy for the jammer in the following restricted sense. The communicator, aware of the jammer's possible knowledge of the code he is going to use, dare not select one that could be rendered useless; therefore, he rules out codeword types with $\Lambda_0(P) \leq \Lambda$. Now the jammer no longer needs to know the actual code to prevent reliable transmission with rate larger than the capacity $C(\Gamma, \Lambda)$; he can always use an independent and identically distributed random state sequence with distribution Q^*.

We emphasize the assumption implicit in our model that the code has to be chosen first and that the jamming strategy may depend on it (though not on the actually transmitted codeword). Of course, there are many different models that could represent practical communication situations. A simple one would require the jamming strategy to be selected first and revealed to both sender and receiver; then the jammer's best strategy would always be the same as in case 1), and the capacity would be equal to $C_r(\Gamma, \Lambda)$. Among the more challenging possibilities is the model solved by Ahlswede [3] where the actual state sequence (rather than just the jamming strategy) is revealed to the sender but not the receiver. We mention that revealing the state sequence to the receiver never leads to a new mathematical problem on account of the possible reduction to the case without such side information (cf. Remark 2 following Theorem 3).

McEliece [14] also considers the case where the commutator and jammer are allowed to use "n-dimensional strategies" $\boldsymbol{X} = (X_1, \cdots, X_n)$ and $\boldsymbol{S} = (S_1, \cdots, S_n)$ satisfying $\mathrm{E}g(\boldsymbol{X}) \leq \Gamma$, $El(\boldsymbol{S}) \leq \Lambda$ (in our notation). By his Theorem 2.1, if the payoff is $(1/n)I(\boldsymbol{X} \wedge \boldsymbol{Y})$ where the output sequence $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ satisfies $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{S}} = W^n$, the value of this game is the same as (6.14) and a pair of optimal strategies consists of independent and identically distributed sequences $\boldsymbol{X}$ and $\boldsymbol{S}$ with distributions yielding a saddle point for (6.14). In seeking a coding interpretation of this result it must be remembered that even for discrete memoryless channels the mutual information between length-n sequences need not be a possible rate of codes with small error probability for the same block length n; rather, as a rule a much larger block length is necessary. This suggests the following modification of model 2) appearing in Theorem 10. Let the random state sequence $\boldsymbol{S} = (S_1, \cdots, S_n)$ be required to consist of independent blocks $(S_1, \cdots, S_k), (S_{k+1}, \cdots, S_{2k}), \cdots, (S_{(l-1)k+1}, \cdots, S_{lk})$ (where $n = lk$, k is fixed), the random variables within each block being allowed to be arbitrarily correlated; further, let each block $\boldsymbol{S}_i = (S_{(i-1)k+1}, \cdots, S_{ik})$ satisfy the constraint $El(\boldsymbol{S}_i) \leq \Lambda$. For this case, Theorem 10 provides the following.

Corollary: The capacity under input constraint Γ and random state selection as above equals $C^{(2)}(\Gamma, \Lambda)$.

Proof: Apply Theorem 10 to the k-block extension W^k of the given AVC, defined by

$$W^k(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}) = \prod_{i=1}^{k} W(y_i|x_i, s_i) \tag{6.15}$$

with the obvious definitions of the cost functions appearing in the input and state constraints. By McEliece's theorem just cited, the random code capacity under input constraint Γ and state constraint Λ of this extended AVC is equal to $kC_r(\Gamma, \Lambda)$. It remains to verify that $\Lambda_0^{(k)}$, the analog for W^k of Λ_0 in fact equals Λ_0.

Let $\mathscr{U}^{(k)}$ denote the family of channels $U^{(k)}$: $\mathscr{X}^k \to \mathscr{S}^k$ for which

$$\sum_{\boldsymbol{s} \in \mathscr{S}^k} W^k(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s}) U^{(k)}(\boldsymbol{s}|\boldsymbol{x}') = \sum_{\boldsymbol{s} \in \mathscr{S}^k} W^k(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{s}) U(\boldsymbol{s}|\boldsymbol{x}). \tag{6.16}$$

Then by (1.6) and (1.4)

$$\begin{aligned} \Lambda_0^{(k)} &= \min_{U^{(k)} \in \mathscr{U}^{(k)}} \ \max_{\boldsymbol{x} \in \mathscr{X}^k} \sum_{\boldsymbol{s} \in \mathscr{S}^k} U^{(k)}(\boldsymbol{s}|\boldsymbol{x}) l(\boldsymbol{s}) \\ &= \frac{1}{k} \min_{U^{(k)} \in \mathscr{U}^{(k)}} \ \max_{\boldsymbol{x} \in \mathscr{X}^k} \sum_{i=1}^{k} \sum_{s \in \mathscr{S}} U_{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k}(s|x_i) l(s) \end{aligned} \tag{6.17}$$

where

$$\begin{aligned} U_{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k}(s|x) &= \sum_{s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_k} U(s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_k | \\ &\qquad x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_k). \end{aligned} \tag{6.18}$$

Since (6.15) and (6.16) imply that the channel (6.18) belongs to $\mathscr{U}$ defined by (1.2) (for every fixed $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k$), (6.17) and (1.6) give $\Lambda_0^{(k)} \geq \Lambda_0$. On the other hand, since the k-block extension U^k of any channel $U \in \mathscr{U}$ obviously belongs to $\mathscr{U}^{(k)}$, and for $U^{(k)} = U^k$ the channel (6.18) equals U, the reverse inequality $\Lambda_0^{(k)} \leq \Lambda_0$ also follows from (6.17) and (1.6).

VII. Conclusion

The general results of Csiszár–Narayan [9] on the capacity for deterministic codes and average probability of error have been applied to specific classes of arbitrarily varying channels, typically with state constraints. For OR channels and group adder channels, explicit or nearly explicit capacity formulae have been obtained. Their noiseless specializations are of independent combinatorial interest, as pointed out here for the OR channel and previously in [9] for two examples which were special group adder channels. For a large class of additive AVC's, an intuitively appealing and simpler form of the general capacity formula was derived with the capacity being positive if and only if the input constraint were less restrictive than the state constraint. For all these classes of AVC's we have shown, using general sufficient conditions derived in Section II, that capacity is attainable using relatively simple decoding rules such as minimum distance for the OR channels and independence (or error vector and codeword) for additive AVC's. We have also discussed how randomized state selection affects capacity, the random state sequence being subject to an expectation constraint. As the expectation constraint alone fails to raise the capacity above the unconstrained one, we have considered various kinds of additional restrictions, each involving independence of the state random variables. In some, but not in all, cases the resulting capacity equals the random code capacity which is also the value of the mutual information game proposed by McEliece [14].

Attention has been restricted in this paper to the discrete case. Continuous alphabet AVC's could be treated with the aid of suitable discrete approximations. This, however, requires mathematical techniques of a different hue and, therefore, will be done elsewhere. We mention, though, that our interest in additive AVC's has largely been motivated by the continuous-alphabet AVC whose output equals the sum of its input, state, and arbitrarily varying noise of variance not exceeding σ^2. Indeed, using the results of Section V (and also the known formula for the value of the corresponding mutual information game) it can be shown that the capacity of this AVC under mean-square input constraint Γ and mean-square state constraint Λ is equal to that of a memoryless channel with signal power Γ and noise power $\Lambda + \sigma^2$ if $\Gamma > \Lambda$, while otherwise the capacity is zero.

Appendix I

Example 1: Let $\mathscr{G}$ be a finite, noncommutative group. Let $\mathscr{X} = \mathscr{S} = \mathscr{Y} = \mathscr{G}$, and let the input x and state s uniquely determine the output as $y = xs$. This deterministic AVC has capacity (even random code capacity) equal to zero; (1.2) is satisfied by $U(s|x) = \text{constant} = 1/|\mathscr{G}|$. However, this AVC is not deterministically symmetrizable. Indeed, deterministic symmetrizability would require the existence of a function f: $\mathscr{G} \to \mathscr{G}$ such that

$$xf(x') = x'f(x) \tag{I.1}$$

for every x and x' in $\mathscr{G}$. Substituting $x = e$ (the identity element of $\mathscr{G}$), we see that f must have the special form $f(x) = xa$, with $a = f(e)$. However, (I.1) then yields $xx'a = x'xa$, or, $xx' = x'x$, which contradicts the assumed noncommutativity of $\mathscr{G}$.

One sufficient condition for an (unconstrained) AVC to have positive capacity is the existence of a distribution P on $\mathscr{X}$ satisfying Condition DS (cf. Definition 3). This condition is due to Dobrushin–Stambler [10]. Another sufficient condition, due to Ahlswede [1], is as follows.

Condition A: There exist distributions P_1 and P_2 on $\mathscr{X}$ such that for no pair of distributions Q_1, Q_3 on $\mathscr{S}$ does

$$\sum_{x,s} P_1(x) Q_1(s) W(y|x,s) = \sum_{x,s} P_2(x) Q_2(s) W(y|x,s) \tag{I.2}$$

hold for every $y \in \mathscr{Y}$.

We now show that Condition A implies Condition DS for some P which in turn implies nonsymmetrizability, but nonsymmetrizability does not imply the existence of any P satisfying DS. This means that, while "DS for some P" and A are sufficient conditions for $C > 0$, neither is necessary.

Notice first that if P does not satisfy DS and P' is arbitrary, then multiplying both sides of (2.9) by $P'(x')$ and summing over x' yields

$$\sum_{x,s} P(x) W(y|x,s) Q'(s) = \sum_{x',s} P'(x') W(y|x',s) Q(s)$$

with $Q'(s) = \Sigma_{x'} U(s|x') P'(x')$. This means that (I.2) holds for $P_1 = P, P_2 = P', Q_1 = Q', Q_2 = Q$. Therefore, if Condition A is satisfied, the distributions P_1 and P_2 therein must satisfy DS. Next, if the AVC is symmetrizable, then multiplying both sides of (1.2) by $P(x)$ and summing over x yields (2.9) with $Q(s) = \Sigma_x U(s|x) P(x)$; therefore, no P can satisfy Condition DS. The next example exhibits a nonsymmetrizable AVC for which no input distribution satisfies DS.

Example 2: Let $\mathscr{X} = \{0,1,2\}$, $\mathscr{S} = \{0,1\}$, $\mathscr{Y} = \{0,1\}$, and define W by

$$W(0|0,0) = 1 \qquad W(0|1,0) = 0.9 \qquad W(0|2,0) = 0.8$$
$$W(0|0,1) = 0 \qquad W(0|1,1) = 0.2 \qquad W(0|2,1) = 0.1.$$

To show that this AVC is nonsymmetrizable, note that if $U(\cdot|\cdot)$ satisfies (1.2), then as $U(1|x) = 1 - U(0|x)$, we obtain from (1.2) for $y = 0$ that

$$W(0|x,1) + [W(0|x,0) - W(0|x,1)] U(0|x')$$
$$= W(0|x',1) + [W(0|x',0) - W(0|x',1)] U(0|x).$$

Substituting $x = 0$, $x' = 1$, followed by $x = 0$, $x' = 2$, and finally, $x = 1$, $x' = 2$, we obtain for the unknowns $U(0|0) = u, U(0|1) = v, U(0|2) = w$, the following three equations:

$$v = 0.2 + 0.7u \qquad w = 0.1 + 0.7u \qquad 0.2 + 0.7w = 0.1 + 0.7v.$$

The inconsistency of this system of equations proves nonsymmetrizability. To show that no input distribution satisfies DS, it suffices to consider (2.9) for $y = 0$ only. Notice that for any distribution P on $\mathscr{X} = \{0,1,2\}$, we have

$$\sum_{x=0}^{2} P(x) W(0|x,0) \geq 0.8 \qquad \sum_{x=0}^{2} P(x) W(0|x,1) \leq 0.2.$$

Taking, for instance, $Q = (1/2, 1/2)$, we observe that

$$0.45 \leq \sum_{s=0}^{1} W(0|x',s) Q(s) \leq 0.55, \qquad \text{for } x' = 0,1,2.$$

Hence it is clear that there exists $U(\cdot|\cdot)$ satisfying (2.9), i.e.,

$$\left[\sum_{x=0}^{2} P(x)W(0|x,0)\right]U(0|x')+\left[\sum_{x=0}^{2} P(x)W(0|x,1)\right]U(1|x') = \sum_{s=0}^{1} W(0|x',s)Q(s)$$

for $x'=0,1,2$, as claimed.

Remarks: 1) Since for AVC's with a binary input alphabet, nonsymmetrizability obviously implies Condition A (set P_1 and P_2 to be point masses at the two input symbols), Example 2 is the simplest possible.

2) By the "strong separation lemma" of Ahlswede [1], Condition A is both necessary and sufficient to render $C>0$ for the class of AVC's with the property that for every channel V: $\mathcal{X} \to \mathcal{S}$, there exists a distribution Q on $\mathcal{S}$ such that

$$\sum_{s} V(s|x)W(y|x,s) = \sum_{s} Q(s)W(y|x,s) \tag{I.3}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. A simple and direct proof showing that for such AVC's Condition A is necessary for nonsymmetrizability, is as follows. If Condition A does not hold, then for every P_1 and P_2 there exist Q_1 and Q_2 satisfying (I.2). Denote by $V_1(\cdot|x,x')$ and $V_2(\cdot|x,x')$ any Q_1 and Q_2 corresponding to the point masses at x and x' in the roles of P_1 and P_2. Then (I.2) becomes

$$\sum_{s} V_1(s|x,x')W(y|x,s) = \sum_{s} V_2(s|x,x')W(y|x',s).$$

By assumption (I.3), the left and right sides above equal $\Sigma_s Q_1(s|x')W(y|x,s)$ and $\Sigma_s Q_2(s|x)W(y|x',s)$, respectively, for suitable $Q_1(\cdot|x')$ and $Q_2(\cdot|x)$. This proves symmetrizability since $U(s|x)=(1/2)[Q_1(s|x)+Q(s|x)]$ satisfies (1.2).

3) Let A'' denote the condition that A is satisfied for blocks of length n (rather than for $n=1$), i.e., that for some distributions $\tilde{P}_1$ and $\tilde{P}_2$ on $\mathcal{X}^n$, there exist no distributions $\tilde{Q}_1$ and $\tilde{Q}_2$ on $\mathcal{S}^n$ such that

$$\sum_{x\in\mathcal{X}^n, s\in\mathcal{S}^n} \tilde{P}_1(x)\tilde{Q}_1(s)W^n(y|x,s) = \sum_{x\in\mathcal{X}^n, s\in\mathcal{S}^n} \tilde{P}_2(x)\tilde{Q}_2(s)W^n(y|x,s).$$

It is known (implicit in [1, sec. 7]) that the validity of A'' for some $n \ge 1$ is both necessary and sufficient for positive capacity. In the terminology of multiuser Shannon theory, this is a "product space characterization" of AVC's with positive capacity, while nonsymmetrizability is the equivalent "single-letter characterization" (cf. the comments in [7, p. 259]). Whereas in multiuser theory such characterizations are typically in terms of information measures, it is striking that we now have a "single letterization" of a "product space characterization," neither of which involves information measures.

Appendix II

Proof of Theorem 1: Readers familiar with [9] will easily realize that the proof is implicitly contained in that of [9, lemma 5]. We now describe the modifications needed to make this explicit. The equation numbers appearing below correspond to those of [9], unless stated otherwise.

a) We wish to establish (3.16), with $I(P)$ replaced by $I(P,\Lambda)$, and $\delta = 3\xi$, for any type for which the given decoding rule is (ξ,τ)-admissible (the hypothesis $P(x)>\beta$ of Lemma 5 is no longer required).

b) In (3.17), replace $I(P)$ by $I(P,\Lambda)$ and skip the paragraph following (3.17).

c) Choose $\eta>0$ small enough to ensure that if the joint type of $\boldsymbol{x}_i, \boldsymbol{s}, \boldsymbol{y}$ belongs to $\mathcal{C}_\eta$, defined by (2.4) with $l(\boldsymbol{s}) \le \Lambda$, as required by the state constraint, then $\boldsymbol{y}$ is $(\boldsymbol{x}_i,\tau)$-typical and (2.5) in our Definition 2 of (ξ,τ)-admissibility is satisfied by dummy random variables X,S,Y, representing this joint type. Then if an error is made when the true codeword, state, and received sequence are as above, our (ξ,τ)-admissible decoding rule must have assigned some candidate message $j \ne i$ to $\boldsymbol{y}$ such that for X, X', S, Y representing the joint type of $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{s}, \boldsymbol{y}$, (2.5) and (2.6) in Definition 2 hold, and also $I(X' \wedge Y) \ge I(P,\Lambda) - \zeta$ (the latter by condition a) in Definition 2). Thus if $\mathcal{D}_\eta$ denotes the set of joint distributions $P_{XX'SY}$ satisfying the latter conditions (instead of those in the paragraph following (3.20)), then (3.21) will remain valid.

d) Equation (3.28) now implies the bound $e_{XX'SY}(i,\boldsymbol{s}) \le \exp\{-n(\delta/3)-3\epsilon)\}$, with $\delta = 3\xi$, simply by the condition $I(X' \wedge Y) \ge I(P,\Lambda) - \xi$ and (3.17) (as modified in b)).

Proof of Theorem 3: On the account of the corollary to Theorem 1, it suffices to prove that for every $\xi>0$, there exists $\tau>0$ and $\delta>0$ such that the typicality decoding rule is (ξ,τ)-admissible for those codeword types P' that satisfy

$$\max_{x} |P(x)-P'(x)| < \delta. \tag{II.1}$$

Now let $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N$ be codewords of common type P'. Observe first that the typicality decoding rule assigns, by definition, at least one message i to any τ-typical $\boldsymbol{y} \in \mathcal{Y}^n$. Then the joint type of $\boldsymbol{x}'=\boldsymbol{x}_i$ and $\boldsymbol{y}$ satisfies (2.2) and hence $I(\boldsymbol{x}' \wedge \boldsymbol{y})$ is arbitrarily close to $I(P',W_Q) \ge I(P',\Lambda)$ if τ is sufficiently small. Hence, condition a) of Definition 2 is satisfied.

Turning next to condition b), it suffices to show that if X, X', S, Y are random variables such that $P_X = P_{X'} = P'$, and

$$\max_{x,y}|P_{X'Y}(x,y) - P'(x)W_Q(y|x)| \le \tau.$$

$$\max_{x,s,y}|P_{XSY}(x,s,y) - P'(x)\tilde{Q}(s)W(y|x,s)| \le \tau,$$

for some Q and $\tilde{Q}$ with $l(Q) \le \Lambda$, $l(\tilde{Q}) \le \Lambda$, then (2.6) must hold if δ (in (II.1)) and τ are sufficiently small.

Suppose indirectly that there exists a sequence of joint distributions $P^{(k)}_{XX'SY}$, $k=1,2,\cdots$, with $P^{(k)}_X = P^{(k)}_{X'} = P^{(k)}$ in the role of P' and satisfying (II.1), where $\delta_k \to 0$, and with

$$\max_{x,y}|P^{(k)}_{X'Y}(x,y) - P^{(k)}(x)W_{Q^{(k)}}(y|x)| \le \tau_k,$$

$$\max_{x,s,y}|P^{(k)}_{XSY}(x,s,y) - P^{(k)}(x)\tilde{Q}^{(k)}(s)W(y|x,s)| \le \tau_k,$$

where $l(Q^{(k)}) \le \Lambda$, $l(\tilde{Q}^{(k)}) \le \Lambda$, and $\tau_k \to 0$, such that neither of these joint distributions meets (2.6) with $\tau = \tau_k$. Then picking a convergent subsequence of $P^{(k)}_{XX'SY}$, its limit $P_{XX'SY}$ will satisfy

$$P_{X'Y}(x',y) = P(x')W_Q(y|x'),$$

$$P_{XSY}(x,s,y) = P(x)\tilde{Q}(s)W(y|x,s) \tag{II.2}$$

for certain Q and $\tilde{Q}$ with $l(Q) \le \Lambda$, $l(\tilde{Q}) \le \Lambda$, and also

$$I(XY \wedge X'|S) = 0.$$

The latter means that

$$P_{XX'SY}(x,x's,y) = P_{X'S}(x',s)P_{XY|S}(x,y|s).$$

Dividing both sides by $P(x')$ and summing over x and s, this results in

$$P_{Y|X'}(y|x') = \sum_{s\in\mathcal{S}} P_{S|X'}(s|x')P_{Y|S}(y|s). \tag{II.3}$$

However, (II.2) and (II.3) mean that P, Q, and $U = P_{S|X'}$ satisfy (2.9) and (2.10), the latter since $l(Q) \le \Lambda$, and

$$\sum_{x,s} P(x)U(s|x)l(s) = \sum_{s} \tilde{Q}(s)l(s) = l(\tilde{Q}) \le \Lambda.$$

This contradiction of the hypothesis on P completes the proof.

Proof of Lemma 1: (1) For an additive AVC defined by (2.13), (1.1) becomes

$$W_Q(y|x) = V_Q(y-x) \qquad \text{with } V_Q(z) = \sum_{s} Q(s)V(z|s). \tag{II.4}$$

With (II.4), the $(\boldsymbol{x}, \tau)$-typicality (2.2) means that

$$\max_{x,y} \left| P_{\boldsymbol{x},\boldsymbol{y}}(x,y) - P(x)V_Q(y-x) \right| \le \tau. \tag{II.5}$$

Substituting $y - x = z$, this provides that the joint type of $\boldsymbol{x}$ and $\boldsymbol{y} - \boldsymbol{x}$ is arbitrarily close to the product distribution $P \times V_Q$; thus (2.14) certainly holds if τ is sufficiently small (depending only on η and the cardinalities of $\mathscr{X}$ and $\mathscr{Y}$).

2) If $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy (2.14), the joint type of $\boldsymbol{x}$ and $\boldsymbol{y} - \boldsymbol{x}$ is close to $P \times T$, where $T = P_{\boldsymbol{y}-\boldsymbol{x}}$. Thus (2.14) implies that

$$\max_{x,y} \left| P_{\boldsymbol{x},\boldsymbol{y}}(x,y) - P(x)T(y-x) \right| \le f(\eta) \tag{II.6}$$

where $f(\eta) \to 0$ as $\eta \to 0$, uniformly in P and T. If $\boldsymbol{y}$ is η-typical, i.e., if it is $(\boldsymbol{x}', \eta)$-typical for some $\boldsymbol{x}'$, then (II.5) also holds with $\boldsymbol{x}'$ instead of $\boldsymbol{x}$ and η instead of τ. From this and (II.6), we obtain by summing over x that

$$\max_{y} \left| (P * V_Q)(y) - (P * T)(y) \right| \le (\eta + f(\eta))|\mathscr{X}|. \tag{II.7}$$

For sufficiently small η, this must imply that

$$\max_{z} \left| V_Q(z) - T(z) \right| < \tau/2, \tag{II.8}$$

say. Indeed, else one could not find P_n, V_{Q_n}, T_n, $n = 1, 2, \cdots$, satisfying (II.7) with $\eta_n \to 0$, such that (II.8) would not hold for V_{Q_n} and T_n, for any n. Then picking a subsequence with $P_{n_k} \to P$, $V_{Q_{n_k}} \to V_Q$, $T_{n_k} \to T$, say, we would get $P * V_Q = P * T$, $V_Q \ne T$, contradicting our hypothesis on $\mathscr{P}$.

Finally, if η is so small that (II.7) implies (II.8) and also $f(\eta) \le \tau/2$, then (II.6) and (II.8) give (II.5), i.e., that $\boldsymbol{y}$ is $(\boldsymbol{x}, \tau)$-typical.

Appendix III

We first show that inequality (3.7) does not hold if Λ is sufficiently close to 1. By Theorem 6, we will then prove that $C(\Lambda) < C_r(\Lambda)$. Set

$$F(\Lambda) = (1-\Lambda)\log\frac{1-r-(1-\Lambda)^2(1-2r)}{r+(1-\Lambda)^2(1-2r)} - \frac{h(r+(1-\Lambda)(1-2r)) - h(r)}{1-2r}. \tag{III.1}$$

We claim that $F(\Lambda) > 0$ if Λ is sufficiently close to 1 (but not equal to 1). Since $h'(t) = \log(1-t)/t$, differentiation yields

$$\begin{aligned} F'(\Lambda) = &-\log\frac{1-r-(1-\Lambda)^2(1-2r)}{r+(1-\lambda)^2(1-2r)} \\ &+(1-\Lambda)\left[\frac{2(1-\Lambda)(1-2r)}{1-r-(1-\Lambda)^2(1-2r)}\right. \\ &\left.+\frac{2(1-\Lambda)(1-2r)}{r+(1-\Lambda)^2(1-2r)}\right] \\ &+\log\frac{1-r-(1-\Lambda)(1-2r)}{r+(1-\Lambda)(1-2r)}. \end{aligned} \tag{III.2}$$

Consider the first case $r \ne 0$. Then $F(1) = F'(1) = 0$, and it is clear from (III.2) that

$$\begin{aligned} F''(1) &= \frac{\partial}{\partial\Lambda}\log\frac{1-r-(1-\Lambda)(1-2r)}{r+(1-\Lambda)(1-2r)}\bigg|_{\Lambda=1} \\ &= \left(\frac{1}{1-r}+\frac{1}{r}\right)(1-2r) > 0 \end{aligned}$$

thereby establishing our claim.

In the case $r = 0$, a stronger result can be proved, namely, that the equation $F(\Lambda) = 0$ has a unique solution $\Lambda^* = 0.6086$ in the interval $(0,1)$, and $F(\Lambda) > 0$ if and only if $\Lambda > \Lambda^*$. In fact, substituting $r = 0$ in (III.1), we get

$$\begin{aligned} F(\Lambda) &= (1-\Lambda)\log\frac{1-(1-\Lambda)^2}{(1-\Lambda)^2} - h(\Lambda) \\ &= (1-\Lambda)\log[(2-\Lambda)\Lambda] - 2(1-\Lambda)\log(1-\Lambda) \\ &\quad + (1-\Lambda)\log(1-\Lambda) + \Lambda\log\Lambda \\ &= \log\Lambda + (1-\Lambda)\log\frac{2-\Lambda}{1-\Lambda} \end{aligned}$$

whence

$$F'(\Lambda) = \frac{1}{\Lambda} - \log\frac{2-\Lambda}{1-\Lambda} - \frac{1-\Lambda}{2-\Lambda} + 1.$$

Now, $F(0) = -\infty$, $F(1) = 0$, $F'(1) = +\infty$, and $F'(\Lambda)$ can be further written as

$$F'(\Lambda) = \frac{2}{\Lambda(2-\Lambda)} - \log\left(1 + \frac{1}{1-\Lambda}\right).$$

Thus $F'(\Lambda)$ is a decreasing function in $(0,1)$ and $F(\Lambda)$ is concave. This proves that the equation $F(\Lambda) = 0$ has a unique solution Λ^* in $(0,1)$, and $F(\Lambda) > 0$ if and only if $\Lambda > \Lambda^*$. Numerical solution of the equation yields $\Lambda^* = 0.6086$.

Next, we prove (3.11) for the noiseless OR channel. For $r = 0$, (3.3) gives

$$I(p,\Lambda) = h((1-p)(1-\Lambda)) - (1-p)h(\Lambda)$$

and

$$\frac{\partial}{\partial p}I(p,\Lambda) = -(1-\Lambda)\log\frac{1-(1-p)(1-\Lambda)}{(1-p)(1-\Lambda)} + h(\Lambda).$$

Solving the equation $(\partial/\partial p)I(p,\Lambda) = 0$ for p, we obtain that $I(p,\Lambda)$ is maximized for

$$p^* = 1 - \frac{1}{1-\Lambda}\left[1 + \exp\frac{h(\Lambda)}{1-\Lambda}\right]^{-1}.$$

It follows, with the notation $u = \exp h(\Lambda)/(1-\Lambda)$, that

$$\begin{aligned} C_r(\Lambda) = I(p^*, \Lambda) &= h\left(\frac{1}{1+u}\right) - \frac{1}{1-\Lambda}\cdot\frac{1}{1+u}h(\Lambda) \\ &= \frac{1}{1+u}\log(1+u) + \frac{u}{1+u}\log\frac{1+u}{u} - \frac{1}{1+u}\log u \\ &= \log(1+u) - \log u = \log\left(1+\frac{1}{u}\right). \end{aligned}$$

Since $1/u = \exp[-h(\Lambda)/(1-\Lambda)] = \exp[\log(1-\Lambda) + (\Lambda/(1-\Lambda))\log\Lambda] = (1-\Lambda)\Lambda^{\Lambda/1-\Lambda}$, this proves (3.11).

References

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

[2] ——, "A method of coding and an application to arbitrarily varying channels," *J. Combin., Inform., Syst. Sci.*, vol. 5, no. 1, pp. 10–35, 1980.

[3] ——, "Arbitrarily varying channels with states sequence known to the sender," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 621–629, Sept. 1986.

[4] R. Ahlswede and J. Wolfowitz, "The capacity of a channel with arbitrarily varying channel probability functions and binary output alphabet," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 15, pp. 186–194, 1970.

[5] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–12, Jan. 1981.

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

[7] ——, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[8] 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.

[9] ——, "The capacity of the arbitrarily varying channel revisited: Capacity, constraints," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 181–193, Mar. 1988.

[10] 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).

[11] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.

[12] C. Heegard and A. A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 731–739, Sept. 1983.

[13] A. V. Kuznetsov and B. S. Tsybakov, "Coding in a memory with defective cells," *Prob. Peredach. Inform.*, vol. 10, no. 2, pp. 52–60, Apr.–June 1974 (English translation).

[14] R. J. McEliece, "Communication in the presence of jamming—An information theory approach," in *Secure Digital Communications*, CISM Courses and Lectures, no. 279, G. Longo, Ed. New York: Springer-Verlag, 1983.

[15] S. Z. Stambler, "Shannon theorems for a full class of channels whose state is known at the output," *Probl. Peredach. Inform.*, vol. 11, no. 4, pp. 263–270, Oct.–Dec. 1975 (English translation).