Channel Capacity for a Given Decoding Metric

Imre Csiszár, *Fellow, IEEE,* and Prakash Narayan, *Senior Member, IEEE*

***Abstract*— For discrete memoryless channels $\{W: \mathcal{X} \to \mathcal{Y}\}$, we consider decoders, possibly suboptimal, which minimize a metric defined additively by a given function $d(x, y) \geq 0$. The largest rate achievable by codes with such a decoder is called the d-capacity $C_d(W)$. The choice $d(x, y) = 0$ if and only if (iff) $W(y|x) > 0$ makes $C_d(W)$ equal to the "zero undetected error" or "erasures-only" capacity $C_{eo}(W)$. The graph-theoretic concepts of Shannon capacity and Sperner capacity are also special cases of d-capacity, viz. for a noiseless channel with a suitable $\{0, 1\}$-valued function d.**

We show that the lower bound on d-capacity given previously by Csiszár and Körner and Hui, is not tight in general, but $C_d(W) > 0$ iff this bound is positive. The "product space" improvement of the lower bound is considered, and a "product space characterization" of $C_{eo}(W)$ is obtained. We also determine the erasures-only (e.o.) capacity of a deterministic arbitrarily varying channel defined by a bipartite graph, and show that it equals capacity. We conclude with a list of challenging open problems.

***Index Terms*— d-decoder, d-capacity, mismatch, Shannon capacity, Sperner capacity, erasures-only capacity, arbitrarily varying channel.**

I. Introduction

In THE traditional Shannon theory of block coding for channels, the primary emphasis has been on the selection of the codeword set. It has been understood that the codeword set will be used in conjunction with an optimal or near-optimal decoder, such as the theoretically optimal maximum-likelihood (ML) decoder or the mathematically convenient joint typicality decoder (cf. standard references such as Cover and Thomas [7], Gallager [18], Wolfowitz [33]). Studying the performance of alternative decoders has not been a major theoretical concern, in contrast with the practical concern of finding computationally feasible or implementable decoding algorithms.

Whereas the ML and joint typicality decoders require a knowledge of the channel, there are several communication situations where the decoder must be designed without such information. The point of view of universal coding has rendered the decoder more theoretically interesting. Indeed, asymptotically optimal decoders can be designed even in the absence of a knowledge of the channel, such as the maximum mutual information decoder for the class of discrete memoryless channels (DMC's) (Goppa [21], Csiszár and Körner [9]); universal decoders for certain channels with memory have been suggested by Ziv [34]. The study of arbitrarily varying channels (AVC's), introduced by Blackwell, Breiman, and Thomasian [6], has placed further emphasis on the issue of decoding. In order to prove AVC coding theorems, quite complex decoding rules had to be devised (Ahlswede [2], Csiszár and Körner [11], Csiszár and Narayan [12]).

In this paper, we shall address the rate of transmission which is attainable on a given channel when the decoding rule is specified, perhaps suboptimally. We concentrate on decoders, termed *d-decoders*, which accept the codeword $\boldsymbol{x}$ "closest" to the received sequence $\boldsymbol{y}$ in the sense of a metric $d(\boldsymbol{x}, \boldsymbol{y})$, defined for sequences as an additive extension of a single-letter metric; the term "metric" is used here in a broad sense as any nonnegative-valued function on the Cartesian product of the input and output alphabets. The optimal rate of transmission achievable on a channel by codes with d-decoding will be termed the *d-capacity* of the channel. A general class of decoders, called α-decoders, has been studied by Csiszár and Körner [10], wherein the metric is an arbitrary function of the joint type of $\boldsymbol{x}$ and $\boldsymbol{y}$ which is not necessarily additive. A more general class of decoders based on pairwise comparisons of codewords relying on the joint types of triples $(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y})$, was introduced in [8] under the name of β-decoders. In [10] and [8], universally attainable exponential error bounds were derived. The lower bound in [10] on the optimal rate attainable with a given α-decoder, specialized to d-decoders, will be one of the starting points of this paper; this lower bound will be denoted by $C_d^{(1)}(W)$. Our focus is on d-decoders, rather than the more general α- or β-decoders, for several reasons:

i) The class of d-decoders itself affords many interesting problems, some of which appear to be very hard.

ii) The study of d-capacity may further enhance the interplay of information theory and combinatorics. Indeed, the important graph-theoretical concepts of Shannon capacity (Shannon [29]) and Sperner capacity (Gargano, Körner, and Vaccaro [20]) are special cases of d-capacity.

iii) As a practical matter, the more general α- and β-decoders, though indispensable for theoretical studies, appear too complex to be implementable. In fact, considerations of complexity may provide a primary reason for the use of a suboptimal decoder.

The study of d-decoders, to our knowledge, was initiated by Stiglitz [31]. He gave an exponential error bound for random

Manuscript received October 29, 1993; revised July 25, 1994. This paper was presented at the Swedish–Russian Information Theory Workshop, Mölle, Sweden, August 1993, and at the IEEE International Symposium on Information Theory, Trondheim, Norway, June 27–July 1, 1994. The work of I. Csiszár was supported by the Hungarian National Foundation for Scientific Research under Grant 1906. The work of P. Narayan was supported by the Institute of Systems Research at the University of Maryland, College Park, under NSF Grant OIR-85-00108, and by the U.S. National Research Council and the Hungarian Academy of Sciences as part of a joint exchange program.

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 Institute of Systems Research, University of Maryland, College Park, MD 20742 USA.

IEEE Log Number 9407439.

codes with an arbitrary decoding metric using Gallager's bounding technique [19]. Fischer [17] gave a lower bound to the best rate attainable on a given DMC when the ML decoder of another channel was used. The lower bound $C_d^{(1)}(W)$ on d-capacity has been derived independently of [10] also by Hui [22], who conjectured the bound to be tight. The use of d-decoders can also be found in the context of spread-spectrum communications (cf. Simon *et al.* [30]). One of the present authors raised the problem of determining the d-capacity of a DMC and some related questions concerning d-decoders at the 1989 Swedish–Soviet Information Theory Workshop [14]. Very recently, several papers have been devoted to this problem area; some of them have been brought to our attention by S. Shamai, to whom we are much indebted. Those prior to ours are Balakirsky [4], Kaplan and Shamai [23] and Merhav, Kaplan, Lapidoth, and Shamai [27]. Overlaps with these in the first version of our paper have been deleted. Also brought to our attention has been the work of Lapidoth [24], showing the lack of tightness of the analog of our lower bound $C_d^{(1)}(W)$ for vector Gaussian channels with Euclidean distance decoding.

Most recently, we have learned of the work of Balakirsky [5] where he has proved the tightness of the bound $C_d^{(1)}(W)$ for channels with binary inputs (announced without proof in [4]), and that of Lapidoth [25] announcing results on d-decoders for multiple-access channels and, as a corollary, a negative solution to *Open Problem 2* raised in this paper. We have also been apprised of the recent independent work of Ahlswede, Cai, and Zhang [3] and Telatar and Gallager [32]. Their results overlap with those in the present paper pertinent to what we term "erasures-only" (e.o.) capacity.

This paper is organized as follows. Section II contains the basic definitions and provides several examples highlighting the scope of the concept of d-capacity. A simple but useful theorem is also included which generalizes a result of Pinsker and Sheverdjaev [28] on e.o. capacity. In Section III, we commence with the lower bound on d-capacity due to Csiszár and Körner [10], and proceed to show that it is not tight in general, but that its positivity is necessary and sufficient for positive d-capacity. We also address the question of the tightness of the "product space" improvement of the lower bound, and obtain a product space characterization of e.o. capacity. In Section IV, we ask if the capacity of deterministic arbitrarily varying channels can be attained by the most elementary decoding rule. Although a complete solution remains elusive, we offer an affirmative answer for a class of deterministic AVC's determined by bipartite graphs. Finally, Section V is devoted to a discussion of several open problems.

We adopt the terminology and notation of the book [9] throughout the paper.

II. Problem Statement and Examples

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets and $d(x, y)$ a nonnegative-valued function on $\mathcal{X} \times \mathcal{Y}$, with $+\infty$ being a possible value of d. For sequences $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ and $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$, we set

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} d(x_i, y_i). \tag{1}$$

For channels with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, we shall consider codes with a decoding metric d, or d-decoding. Such a code is defined by a codeword set $\{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\} \subset \mathcal{X}^n$ and a decoder which assigns to a received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ that message i for which $d(\boldsymbol{x}^{(i)}, \boldsymbol{y}) < d(\boldsymbol{x}^{(j)}, \boldsymbol{y})$ for all $j \neq i$; if for no such i exists, an error is declared.

Definition 1: The d-capacity of a channel is the supremum of those numbers R for which, for every $\epsilon > 0$ and sufficiently large n, there exist codes with d-decoding such that the rate is $(1/n) \log N > R$ and the (average or maximum) probability of error is less than ϵ.

Clearly, the d-capacity does not depend on whether the average or maximum error criterion is used.

In this paper, we shall be concerned mainly with discrete memoryless channels (DMC's). The d-capacity of a DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ will be denoted by $C_d(W)$.

Examples: 1) For $d(x, y) = -\log W(y|x)$, d-decoding is the same as strict maximum-likelihood (ML) decoding, i.e., the message i is accepted if and only if (iff) $W(\boldsymbol{y}|\boldsymbol{x}^{(i)}) > W(\boldsymbol{y}|\boldsymbol{x}^{(j)})$ for all $j \neq i$. It is well-known that the capacity $C(W)$ of a DMC $\{W\}$ can be attained by using strict ML decoding. Thus in this case, $C_d(W) = C(W)$.

2) Sometimes a decoding metric $d(x, y) = -\log V(y|x)$ is used, where V is a channel different from the true channel W. (This may happen, for instance, when the true channel is unknown to the receiver.) This situation is often referred to as mismatched decoding; in this case, $C_d(W) \leq C(W)$, and the exact value of $C_d(W)$ is not known in general.

Of particular interest to use are those decoding metrics d for which $d(x, y) = 0$ whenever $W(y|x) > 0$. Then, the d-decoder accepts message i iff it is the only message with $d(\boldsymbol{x}^{(i)}, \boldsymbol{y}) = 0$; if more than one such message exist, an error is declared. Thus an incorrect message is never accepted; the only errors are erasures. The following are special cases of such decoding metrics, which we shall refer to as *erasures-only (e.o.) metrics*.

3) Let $d(x, y) = 0$ iff $W(y|x) > 0$. This metric results in the smallest possible probability of erasure while permitting no undetected errors. Thus in this case $C_d(W)$ equals the so-called "zero error capacity with erasures" or "zero undetected error capacity" of the DMC $\{W\}$ (cf. Pinsker and Sheverdjaev [28]). In this paper, it will be referred to as "erasures-only" capacity or e.o. capacity, denoted by $C_{eo}(W)$.

4) The Shannon capacity [29] of a graph $\mathcal{G}$ with vertex set $\mathcal{X}$ is defined as the limit as $n \to \infty$ of $(1/n) \log \alpha(\mathcal{G}^n)$, where $\mathcal{G}^n$ is the graph with vertex set $\mathcal{X}^n$ such that $(\boldsymbol{x}, \boldsymbol{x}')$, $\boldsymbol{x} \neq \boldsymbol{x}'$, is an edge iff for each $1 \leq l \leq n$, either $x_l = x_l'$ or else (x_l, x_l') is an edge of $\mathcal{G}$; and $\alpha(\mathcal{G}^n)$ is the maximum cardinality of a set $\mathcal{C} \subseteq \mathcal{X}^n$ such that no $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\mathcal{C}$ are connected by an edge of $\mathcal{G}^n$. The zero-error capacity of a DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ is equal to the Shannon capacity of the graph with vertex set $\mathcal{X}$ in which x and x', $x \neq x'$, are connected by an edge iff $W(y|x)W(y|x') > 0$ for some $y \in \mathcal{Y}$. Now, the Shannon

capacity of $\mathcal{G}$ is equal to the d-capacity of the noiseless channel with alphabet $\mathcal{X}$, where d is such that $d(x, y) = 0$ iff $x = y$ or (x, y) is an edge of $\mathcal{G}$. Indeed, for the noiseless channel, the maximum probability of error (for any code) is either 0 or 1, and d-decoding with d as chosen above gives zero probability of error iff the codewords are all distinct and no two of them are connected by an edge of $\mathcal{G}^n$.

5) Gargano, Körner, and Vaccaro [20] have defined the Sperner capacity of a directed graph $\mathcal{G}$ with vertex set $\mathcal{X}$ as the limit as $n \to \infty$ of $(1/n) \log N_n(\mathcal{G})$, where $N_n(\mathcal{G})$ is the maximum cardinality of a set $\mathcal{C} \subseteq \mathcal{X}^n$ such that for every $\boldsymbol{x} \neq \boldsymbol{x}'$ in $\mathcal{C}$, there exists $1 \leq l \leq n$ for which (x_l, x_l') is an edge of $\mathcal{G}$. An undirected graph can be identified with a directed graph such that if (x, x') is an edge, then so is (x', x). With this understanding, the Sperner capacity of an undirected graph is equal to the Shannon capacity of its complementary graph. Now, given a directed graph $\mathcal{G}$, let $d(x, y) > 0$ iff (x, y) is an edge of $\mathcal{G}$. Then, by the same argument as in Example 4 above, the d-capacity of the noiseless channel with alphabet $\mathcal{X}$ is equal to the Sperner capacity of $\mathcal{G}$.

Examples 4 and 5 suggest that a general single-letter formula for d-capacity is not to be expected soon.

Remark (cf. also Balakirsky [4]): In Definition 1, attention could have been restricted to codes with codewords of the same type (since any codeword set contains a subset of exponentially the same size, consisting of codewords of identical type). Hence, if d and $\tilde{d}$ are such that

$$d(x, y) = c(a(x) + b(y) + \tilde{d}(x, y)) \tag{2}$$

for some finite numbers $a(x)$, $x \in \mathcal{X}$, $b(y)$, $y \in \mathcal{Y}$, and $c > 0$, then $C_d(W) = C_{\tilde{d}}(W)$ for *every* DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ (since if the codewords $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}$ are of the same type and (2) holds, then by (1) the d-decoder accepts message i iff the $\tilde{d}$-decoder does).

As a special case of the remark above, we get the following sufficient condition for a DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ to have $C_{eo}(W) = C(W)$ (cf. Example 3).

Theorem 1: Suppose that there exist positive numbers $A(x)$, $x \in \mathcal{X}$, $B(y)$, $y \in \mathcal{Y}$, such that

$$W(y|x) = A(x)B(y) \quad \text{whenever } W(y|x) > 0. \tag{3}$$

Then $C_{eo}(W) = C(W)$.

Proof: Let $\tilde{d}(x, y) = -\log W(y|x)$ and

$$d(x, y) = \begin{cases} 0, & \text{if } W(y|x) > 0 \\ +\infty, & \text{if } W(y|x) = 0. \end{cases} \tag{4}$$

Then (2) holds with $a(x) = \log A(x)$, $b(y) = \log B(y)$, so that $C_d(W) = C_{\tilde{d}}(W)$. Here $C_d(W) = C_{eo}(W)$ by Example 3 and $C_{\tilde{d}}(W) = C(W)$ by Example 1.

Remark: Pinsker and Sheverdjaev [28] have proved that $C_{eo}(W)$ equals $C(W)$ if there do not exist distinct elements $x_1, \cdots, x_l$ of $\mathcal{X}$ and distinct elements $y_1, \cdots, y_l$ of $\mathcal{Y}$ such that $W(y_i|x_i) > 0$ and $W(y_i|x_{i+1}) > 0$ for $i = 1, \cdots, l$, where $x_{l+1} = x_1$. It is easily seen that W with this property satisfies the hypothesis of Theorem 1; thus Theorem 1 is a generalization of the Pinsker and Sheverdjaev result [28].

Note that the sufficient condition in Theorem 1 can be weakened in an obvious manner if the channel $\{W: \mathcal{X} \to \mathcal{Y}\}$ has a capacity-achieving distribution concentrated on a subset $\mathcal{X}_0$ of $\mathcal{X}$. Then it suffices that (3) holds when $x \in \mathcal{X}_0$, $y \in \mathcal{Y}$. We conjecture that this weakened condition is also necessary for $C_{eo}(W) = C(W)$.

III. Lower Bound on, and Positivity of $C_d(W)$

A general class of decoding rules, called α-decoding rules, has been considered by Csiszár and Körner in [10]. Given a function α on the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$, the α-decoder accepts message i iff

$$\alpha(P_{\boldsymbol{x}^{(i)}\boldsymbol{y}}) < \alpha(P_{\boldsymbol{x}^{(j)}\boldsymbol{y}}), \qquad \text{for all } j \neq i \tag{5}$$

where $P_{\boldsymbol{x}^{(i)}\boldsymbol{y}}$ denotes the joint type of $\boldsymbol{x}^{(i)} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$. To be exact, this may be termed the strict α-decoder, since in [10] ties were permitted to be broken arbitrarily unlike above. However, the results in [10] obviously hold for our case too with the understanding that if no i satisfying (5) exists, an error is declared.

It is shown in [10] that for every $R > 0$, every n, and every codeword type P, there exist codes of block length n with α-decoder, with codewords of type P and rate $\geq R - \delta_n$, such that for *every* DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ the maximum probability of error is $\leq \exp[-n(E_{\alpha, r}(R, P, W) - \delta_n')]$, where $\delta_n \to 0$ and $\delta_n' \to 0$ are explicitly given sequences (not depending on R, P, W, α), and

$$E_{\alpha, r}(R, P, W) = \inf_{\substack{PV = P\tilde{V} \\ \alpha(P, \tilde{V}) \leq \alpha(P, V)}} (D(V\|W|P) + |I(P, \tilde{V}) - R|^+) \tag{6}$$

the infimum being over all pairs of auxiliary channels V, $\tilde{V}$ satisfying the indicated constraints. Here $\alpha(P, V)$ denotes the value of α at the distribution on $\mathcal{X} \times \mathcal{Y}$ defined by $P(x)V(y|x)$. In [10], the definition (6) of $E_{\alpha, r}(R, P, W)$ is stated with *min* rather than *inf*; this is a slight error since for a general α the minimum need not be attained. If α is sufficiently regular so that the minimum in (6) is attained, the conclusion in [10] is valid that $E_{\alpha, r}(R, P, W) > 0$ iff R is less than the minimum of $I(P, \tilde{V})$ subject to $P\tilde{V} = PW$ and $\alpha(P, \tilde{V}) \leq \alpha(P, W)$.

Clearly, our d-decoder is an α-decoder in the sense above, with $\alpha(P, V) = \Delta(P, V)$ defined by

$$\Delta(P, V) = \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P(x)V(y|x)d(x, y) \tag{7}$$

and for this choice of α, the minimum in (6) is always attained. Indeed, this is trivial (with the minimum being 0) if $d(x, y) = +\infty$ for some (x, y) with $P(x)W(y|x) > 0$; otherwise, we minimize a continuous function of the pair $(V, \tilde{V})$ on the compact set determined by the constraints in (6), with the additional one that $V(y|x) = 0$ whenever $W(y|x) = 0$. Hence, as a simple corollary of the result above in [10], we have the following proposition.

Proposition 1: For any decoding metric d

$$C_d(W) \geq \max_P I_d(P, W) \tag{8}$$

where

$$I_d(P, W) = \min_{\substack{P_X = P,\, P_Y = PW \\ Ed(X, Y) \leq \Delta(P, W)}} I(X \wedge Y). \tag{9}$$

Moreover, for any distribution P on $\mathcal{X}$ and $R > 0$, there exist constant composition codes with d-decoder, with codeword type approaching P, rate approaching R, and probability of error going to zero exponentially for every DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ such that $R < I_d(P, W)$.

Proof: The proof is immediate from the results in [10] stated above, using the continuity of $I_d(P, W)$ (cf. the next Lemma).

Remark: The lower bound in (8) has been obtained, independently of [10], also by Hui [22]. The second part of the following lemma appears in [22] as well.

Lemma 1: $I_d(P, W)$ is a continuous function of the pair (P, W) if d is finite-valued; otherwise, it is continuous when W is restricted to the set of channels such that $W(y|x) = 0$ whenever $d(x, y) = \infty$. Furthermore, $I_d(P, W)$ is positive iff

$$\sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P(x)(PW)(y)d(x, y) > \Delta(P, W) \tag{10}$$

and, if (10) holds, the inequality constraint in (9) can be replaced by equality.

Proof: See Appendix.

Example 6 (also cf. Hui [22], Balakirsky [4]): For a binary channel, i.e., when $|\mathcal{X}| = |\mathcal{Y}| = 2$, the constraints $P_X = P$, $P_Y = PW$, and $Ed(X, Y) = \Delta(P, W)$ in (9) force $P_{XY}(x, y) = P(x)W(y|x)$. Thus by Lemma 1, for a binary channel with arbitrary decoding metric $d(x, y)$, $I_d(P, W)$ is either equal to $I(P, W)$ or 0, according to whether or not (10) holds. Simple algebra shows that for a binary channel, condition (10) reduces to

$$d(0, 1) + d(1, 0) \gtrless d(0, 0) + d(1, 1) \tag{11}$$

accordingly as the sum of the channel crossover probabilities is $\lesseqgtr 1$, independently of the input distribution. It follows that if (11) holds, then $C_d(W) = C(W)$. (This also follows from the remark in Section II, viz. (2).) On the other hand, if (11) fails to hold, then $C_d(W) = 0$ by Theorem 3 below.

Remarks: i) The lower bound of Fischer [17] on d-capacity with $d(x, y) = -\log V(y|x)$ (cf. Example 2) is the maximum of

$$\sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P(x)W(y|x) \log \frac{V(y|x)}{(PV)(y)}$$

with respect to P. It can be seen by simple algebra that if X and Y satisfy the constraints in (9) with $d(x, y) = -\log V(y|x)$, then $I(X \wedge Y)$ is lower bounded by the sum above. Thus

$$I_d(P, W) \geq \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P(x)W(y|x) \log \frac{V(y|s)}{(PV)(y)}, \quad \text{if } d(x, y) = -\log V(y|x). \tag{12}$$

The lower bound in (12) on $I_d(P, W)$ is a special case of more general bounds developed in [23].

ii) Since the codes appearing in Proposition 1 do not depend on the channel $\{W\}$, a sender–receiver pair lacking knowledge of $\{W\}$ other than its membership of a certain class of channels $\mathcal{W}$, can find a good code with d-decoding, of any rate

$$R < \max_P \inf_{W \in \mathcal{W}} I_d(P, W). \tag{13}$$

In other words, the right-hand side above is a lower bound on the d-capacity of the compound channel defined by the set $\mathcal{W}$. Of particular interest is the case in which $\mathcal{W}$ is a convex compact set when the ordinary capacity of the compound channel is

$$C(\mathcal{W}) = \max_P \min_{W \in \mathcal{W}} I(P, W) = \min_{W \in \mathcal{W}} \max_P I(P, \mathcal{W}). \tag{14}$$

Let (P^*, W^*) be the saddle-point in (14), i.e., $C(\mathcal{W}) = I(P^*, W^*)$. Then $C(\mathcal{W})$ can be achieved by d-decoding with $d(x, y) = -\log W^*(y|x)$. Indeed, we have for every $W \in \mathcal{W}$ that

$$\begin{aligned} I_d(P^*, W) &\geq \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P^*(x)W(y|x) \log \frac{W^*(y|x)}{(P^*W^*)(y)} \\ &\geq I(P^*, W^*) = C(\mathcal{W}) \end{aligned} \tag{15}$$

where the first inequality is (12), and the second is a consequence of the fact that W^* minimizes $I(P^*, W)$ over the convex compact set $\mathcal{W}$ (cf. [9, p. 213, eq. (6.19)]). On the other hand, it should be noted that d-decoders are absolutely inadequate for certain compound channels, in that the d-capacity of a compound channel of positive capacity may equal 0 for every decoding metric $d(x, y)$. This happens, for instance, if $\mathcal{W}$ consists of two binary symmetric channels with crossover probabilities $1/4$ and $3/4$ (cf. Example 6).

The lower bound on $C_d(W)$ in Proposition 1 is, in general, not tight, and may be improved by recourse to a product space version. Namely, setting

$$C_d^{(n)}(W) = \frac{1}{n} \max_{\tilde{P}} I_d(\tilde{P}, W^n) \tag{16}$$

where $\tilde{P}$ ranges over all distributions on $\mathcal{X}^n$, it follows from Proposition 1 that $C_d(W) \geq C_d^{(n)}(W)$ for all n, and consequently

$$C_d(W) \geq C_d^{(\infty)}(W) = \sup_n C_d^{(n)}(W). \tag{17}$$

Of course, the "product space" lower bound in (17) is, in general, not computable. We shall return to the issue of the tightness of this bound later in this section. At this point, we show that $C_d^{(n)}(W)$, $n > 1$, may be larger than $C_d^{(1)}(W)$, the bound appearing in (8). Indeed, this is already true in the special case of Example 4 on account of the following Lemma and the fact that $(1/n) \log \alpha(\mathcal{G}^n)$ may be greater than $\alpha(\mathcal{G})$ (e.g., if $\mathcal{G}$ is the pentagon, then $\alpha(\mathcal{G}) = 2$ and $\alpha(\mathcal{G}^2) = 5$).

Lemma 2: Let $\mathcal{G}$ be an undirected graph with vertex set $\mathcal{X}$, and let $d(x, y)$ be positive iff $x \neq y$ and (x, y) is not an edge of $\mathcal{G}$. Then for the noiseless channel W_0 with alphabet $\mathcal{X}$, we have

$$C_d^{(n)}(W_0) = \frac{1}{n} \log \alpha(\mathcal{G}^n).$$

Proof: See Appendix.

Remark: It is possible that $C_d^{(n)}(W) > C_d^{(1)}(W)$ even for $\{W\}$ such that $W(y|x) > 0$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Indeed, let $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3, 4\}$, and $d(x, y) = 0$ iff $y = x - 1$, x or $x + 1$ (mod 5). Consider any sequence of channels $\{W: \mathcal{X} \to \mathcal{Y}\}$ converging to the noiseless channel $\{W_0\}$. Then for any fixed P, $I_d(P, W)$ converges by Lemma 1 to $I_d(P, W_0)$, and the same holds for $I_d(\tilde{P}, W^2)$ too. Since $C_d^{(1)}(W_0) = \log 2 < C_d^{(2)}(W_0) = \frac{1}{2}\log 5$, the inequality $C_d^{(1)}(W) < C_d^{(2)}(W)$ will hold for every $\{W\}$ sufficiently close to $\{W_0\}$.

Even though the lower bound in Proposition 1 is, in general, not tight, we now show that the positivity of that bound is necessary for $C_d(W) > 0$. Indeed, it is necessary even for the distinguishability of two codewords by d-decoding.

Theorem 2: Given a DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ and a decoding metric $d(x, y)$, suppose that for every $\epsilon > 0$ there exist some n and $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{x}' = (x_1', \cdots, x_n')$ in $\mathcal{X}^n$ such that the maximum probability of error of the code with codeword set $\{\boldsymbol{x}, \boldsymbol{x}'\}$ and d-decoding is less than ϵ. Then

$$C_d^{(1)}(W) = \max_P I_d(P, W) > 0.$$

Remark: For achieving positive rates with constant composition codes of codeword type P using d-decoding, it is not necessary that $I_d(P, W)$ be positive. Rather, positive rates can always be achieved by such codes whenever $C_d(W) > 0$, at least if $P(x)$ is bounded away from zero. To see this, take any code with codewords of type P_0 such that d-decoding yields probability of error less than ϵ. The probability of error with d-decoding will remain unchanged if we append to each codeword a fixed sequence; if this appended sequence is suitably selected (with length not exceeding a constant times the original block length), the resulting new code will have codewords of type P, and rate not less than a constant times the original rate.

In order to prove Theorem 2, we need the following lemma.

Lemma 3: Let $\mathcal{P}$ be a finite set of distributions on the real line, each concentrated on a finite set. Then there exists an $\epsilon > 0$ such that for every n and independent random variables $X_1, \cdots, X_n$ with $P_{X_i} \in \mathcal{P}$, $i = 1, \cdots, n$, the condition

$$\Pr\left\{\sum_{i=1}^n X_i \geq 0\right\} < \epsilon \tag{18}$$

implies

$$\sum_{i=1}^n EX_i < 0. \tag{19}$$

Proof: See Appendix.

Proof of Theorem 2: Let $(Y_1, \cdots, Y_n)$ denote the sequence of output random variables resulting when $\boldsymbol{x}$ is sent, i.e., let $Y_1, \cdots, Y_n$ be independent with distributions $P_{Y_i} = W(\cdot|x_i)$, $i = 1, \cdots, n$. Then, the probability of error when $\boldsymbol{x}$ is sent, is

$$P_e(\boldsymbol{x}) = \Pr\left\{\sum_{i=1}^n d(x_i, Y_i) \geq \sum_{i=1}^n d(x_i', Y_i)\right\}. \tag{20}$$

Suppose that $P_e(\boldsymbol{x}) < \epsilon$ with $\epsilon > 0$ sufficiently small. Then each $d(x_i, Y_i)$ must be finite with probability 1; else, $P_e(\boldsymbol{x})$ could not be less than the smallest positive $W(y|x)$. If each $d(x_i', Y_i)$ as well is finite with probability 1, we obtain from (20) and Lemma 3 that

$$\sum_{i=1}^n E[d(x_i, Y_i) - d(x_i', Y_i)] < 0. \tag{21}$$

If $d(x_i', Y_i) = \infty$ with positive probability for some i, then (21) holds trivially.

Similarly, denoting by $(Y_1', \cdots, Y_n')$ the output random variables resulting from the transmission of $\boldsymbol{x}'$, i.e., $Y_1', \cdots, Y_n'$ are independent with $P_{Y_i'} = W(\cdot|x_i')$, $i = 1, \cdots, n$, it follows from $P_e(\boldsymbol{x}') < \epsilon$ that

$$\sum_{i=1}^n E[d(x_i', Y_i') - d(x_i, Y_i')] < 0. \tag{22}$$

Adding (21) and (22), we conclude that the sum of the expectations

$$E[d(x_i, Y_i) + d(x_i', Y_i') - d(x_i', Y_i) - d(x_i, Y_i')]$$

is negative, so that at least one of them is negative. Hence, there exist x and x', $x \neq x'$ in $\mathcal{X}$ such that

$$\sum_{y \in \mathcal{Y}} (d(x, y)W(y|x) + d(x', y)W(y|x')) < \sum_{y \in \mathcal{Y}} (d(x', y)W(y|x) + d(x, y)W(y|x')). \tag{23}$$

Finally, let P be the distribution on $\mathcal{X}$ which assigns probability $\frac{1}{2}$ to both x and x' appearing in (23). It is seen by simple algebra that (23) is equivalent to (10) for this P, so that by Lemma 1 we have $I_d(P, W) > 0$. This completes the proof of Theorem 2.

We now return to the issue of whether or not the product space lower bound (17) is tight. Although this bound is not computable, its tightness would afford some valuable conclusions, for instance, that for $R < C_d(W)$, codes with d-decoding always exist with rates approaching R and probability of error approaching zero exponentially fast. Indeed, this is certainly true for $R < C_d^{(\infty)}(W)$ by Proposition 1. We believe that the bound in (17) is tight for every decoding metric $d(x, y)$, but at present cannot offer a proof. However, it is easy to prove the "product space characterization" of $C_d(W)$ for e.o. metrics.

Theorem 3: For a DMC $\{W\}$, $C_d(W) = C_d^{(\infty)}(W)$ holds for every e.o. metric $d(x, y)$.

Proof: Let $\mathcal{C} \subset \mathcal{X}^n$ be the codeword set of a code with d-decoder, where d is an e.o. metric, i.e., $d(x, y) = 0$ if $W(y|x) > 0$. Let $\mathcal{D}$ denote the set of those $\boldsymbol{y} \in \mathcal{Y}^n$ which are correctly decoded, i.e.,

$$\mathcal{D} = \{\boldsymbol{y}: \text{ there is a unique } \boldsymbol{x} \in \mathcal{C} \text{ with } d(\boldsymbol{x}, \boldsymbol{y}) = 0\}. \tag{24}$$

Then the average probability of error is

$$P_e = \frac{1}{|\mathcal{C}|}\sum_{\boldsymbol{x} \in \mathcal{C}} W^n(\mathcal{D}^c|\boldsymbol{x}) = (\tilde{P}W^n)(\mathcal{D}^c) \tag{25}$$

where $\tilde{P}$ denotes the uniform distribution on $\mathcal{C}$.

Now, $I_d(\tilde{P}, W^n)$ equals, by definition, the minimum of $I(X^n \wedge Y^n)$ subject to the constraints (cf. (9))

$$P_{X^n} = \tilde{P} \qquad P_{Y^n} = \tilde{P}W^n \qquad Ed(X^n, Y^n) = 0. \quad (26)$$

The last condition in (26) yields that $d(X^n, Y^n) = 0$ with probability 1, and hence, X^n is uniquely determined by Y^n if $Y^n \in \mathcal{D}$. Thus

$$H(X^n|Y^n) = \sum_{\boldsymbol{y} \in \mathcal{D}^c} \Pr\{Y^n = \boldsymbol{y}\} H(X^n|Y^n = \boldsymbol{y}) \leq \log|\mathcal{C}| \Pr\{Y^n \in \mathcal{D}^c\}$$

and

$$I(X^n \wedge Y^n) = \log|\mathcal{C}| - H(X^n|Y^n) \geq (1 - \Pr\{Y^n \in \mathcal{D}^c\}) \log|\mathcal{C}|.$$

It follows, using (25), that

$$I_d(\tilde{P}, W^n) \geq (1 - P_e) \log|\mathcal{C}|$$

so that the rate of our code is bounded above according to

$$\frac{1}{n} \log|\mathcal{C}| \leq \frac{1}{1 - P_e} \cdot \frac{1}{n} I_d(\tilde{P}, W^n) \leq \frac{1}{1 - P_e} C_d^{(\infty)}.$$

This proves that $C_d(W) \leq C_d^{(\infty)}$; comparing this with (17) completes the proof of Theorem 3.

IV. "Erasures-Only" Capacity for Arbitrarily Varying Channels

The performance of various decoding rules for AVC's has been investigated by Csiszár and Narayan [13]. In this section, we address the following question: Can the capacity of a *deterministic* AVC be attained using the most elementary decoding rule; namely, a codeword $\boldsymbol{x}$ is accepted iff it is the only codeword compatible with the received sequence $\boldsymbol{y}$?

Formally, given an input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and a set of states $\mathcal{S}$, a deterministic AVC is defined by a family of mappings T_s: $\mathcal{X} \to \mathcal{Y}$, $s \in \mathcal{S}$, where $y = T_s(x)$ is the (only) output for input x if the state is s. During the transmission of a sequence $\boldsymbol{x} = (x_1, \cdots, x_n)$, the states may vary in an arbitrary manner; if the state sequence is $\boldsymbol{s} = (s_1, \cdots, s_n) \in \mathcal{S}^n$, the output will be $\boldsymbol{y} = T_{\boldsymbol{s}}(\boldsymbol{x}) = (T_{s_1}(x_1), \cdots, T_{s_n}(x_n))$.

For a code with codeword set $\mathcal{C} \subset \mathcal{X}^n$ and a given decoder, let $p_e(\boldsymbol{s})$ denote the fraction of codewords $\boldsymbol{x} \in \mathcal{C}$ that are incorrectly decoded when the sequence of states is $\boldsymbol{s} = (s_1, \cdots, s_n)$. We define the capacity of the deterministic AVC as the supremum of those numbers R for which, for any $\epsilon > 0$ and sufficiently large n, there exist codes of rate $(1/n)\log|\mathcal{C}| > R$ with a suitable decoder, such that $p_e(\boldsymbol{s}) < \epsilon$ for *all* $\boldsymbol{s} \in \mathcal{S}^n$. In the terminology of arbitrarily varying channels, this is the capacity for deterministic codes and the average probability of error criterion, when both the encoder and the decoder are ignorant of the actual sequence of states (cf. [12]). Observe that since we are dealing with a family of deterministic channels, the maximum probability of error criterion would require that every codeword be correctly decodable, regardless of the state sequence.

Let us consider the (elementary) decoding rule whereby a codeword $\boldsymbol{x}$ is accepted iff it is the only codeword for which a state sequence $\boldsymbol{s} \in \mathcal{S}^n$ exists such that $T_{\boldsymbol{s}}(\boldsymbol{x})$ equals the received sequence $\boldsymbol{y}$, i.e., the d-decoder with metric $d(x, y)$ such that $d(x, y) = 0$ iff there exists $\boldsymbol{s} \in \mathcal{S}$ with $T_s(x) = y$. For this decoder, we have

$$p_e(\boldsymbol{s}) = \frac{1}{|\mathcal{C}|} |\{\boldsymbol{x} \in \mathcal{C}: \text{there exist } \boldsymbol{x}' \in \mathcal{C}, \boldsymbol{s}' \in \mathcal{S}^n \text{ with } \boldsymbol{x}' \neq \boldsymbol{x}, T_{\boldsymbol{s}'}(\boldsymbol{x}') = T_{\boldsymbol{s}}(\boldsymbol{x})\}|. \quad (27)$$

Definition 2: The e.o. capacity of a deterministic AVC is defined as the supremum of those numbers R for which, for every $\epsilon > 0$ and sufficiently large n, there exist codes of rate $(1/n)\log|\mathcal{C}| > R$ such that $p_e(\boldsymbol{s}) < \epsilon$ for all $\boldsymbol{s} \in \mathcal{S}^n$, where $p_e(\boldsymbol{s})$ is defined by (27).

It is tempting to conjecture that the e.o. capacity of a deterministic AVC always equals its capacity. This, however, is apparently a difficult problem; below we shall establish this equality for a subclass of deterministic AVC's.

Let $\mathcal{G}$ be a bipartite graph with vertex sets $\mathcal{X}$ and $\mathcal{Y}$, with no isolated vertices. We interpret $\mathcal{X}$ and $\mathcal{Y}$ as the input and output alphabets of a channel. At each time instant, each $x \in \mathcal{X}$ is connected to some $y \in \mathcal{Y}$ such that (x, y) is an edge of $\mathcal{G}$; each $x \in \mathcal{X}$ is connected to just a single $y \in \mathcal{Y}$. These connections may change in time in an arbitrary manner. Thus the graph $\mathcal{G}$ defines a deterministic AVC such that i) for each state $s \in \mathcal{S}$, the mapping T_s: $\mathcal{X} \to \mathcal{Y}$ represents a possible pattern of connections, i.e., a deterministic channel such that $(x, T_s(x))$ is an edge of the graph $\mathcal{G}$; and ii) the mappings T_s, $s \in \mathcal{S}$, exhaust all possible patterns of connections as above.

Given the bipartite graph $\mathcal{G}$, we shall denote by $\tilde{\mathcal{G}}$ the graph with vertex set $\mathcal{X}$ for which (x, x') is an edge iff there exists $y \in \mathcal{Y}$ such that both (x, y) and (x', y) are edges of $\mathcal{G}$.

Theorem 4: The e.o. capacity of the deterministic AVC defined by a bipartite graph $\mathcal{G}$ as above is equal to the capacity of this AVC. This capacity is positive iff $\tilde{\mathcal{G}}$ is not the complete graph. If the latter condition holds, the capacity equals the minimum capacity of DMC's $\{W: \mathcal{X} \to \mathcal{Y}\}$ compatible with $\mathcal{G}$, i.e., such that $W(y|x)$ is always zero if (x, y) is not an edge of $\mathcal{G}$.

Remark: Note that the condition for the positivity of the e.o. capacity of a deterministic AVC determined by a bipartite graph $\mathcal{G}$, is the same as that for the zero error capacity of the same AVC, which equals the Shannon capacity of the graph $\tilde{\mathcal{G}}$ [2]. However, the e.o. capacity of this AVC (which equals capacity) itself may be strictly larger than the zero-error capacity. For example, consider the graph $\mathcal{G}$ of Fig. 1. The zero-error capacity equals the Shannon capacity of the pentagon, i.e., $1/2 \log 5$ [26], whereas the capacity is equal to $\log 5 - 1$.

Proof of Theorem 4: We first show that the condition for positive capacity in the theorem is indeed necessary. (Its sufficiency is obvious.) For this, we rely on the fact that symmetrizable AVC's have capacity 0. The simplest, deterministic version of symmetrizability for a general AVC states that there exists a mapping u: $\mathcal{X} \to \mathcal{S}$ such that for every x and x' in $\mathcal{X}$, the output distribution when x is sent and $s' = u(x')$ is

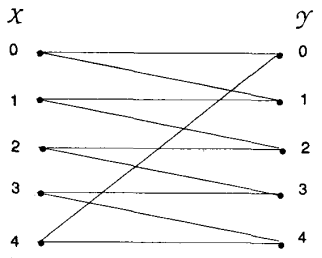

Fig. 1.

the state, is the same as that when x' is sent and $s = u(x)$ is the state. For a deterministic AVC, this means that

$$T_{u(x')}(x) = T_{u(x)}(x'), \qquad \text{for every } x \text{ and } x' \in \mathcal{X}. \quad (28)$$

Now, if $\tilde{\mathcal{G}}$ is the complete graph, assign to each pair (x, x') a $y = y(x, x')$ such that both (x, y) and (x', y) are edges of $\mathcal{G}$, and $y(x, x') = y(x', x)$. Then define the mapping $u: \mathcal{X} \to \mathcal{S}$ as follows: For each $x \in \mathcal{X}$, let $s = u(x)$ be the state for which the (deterministic) channel sends each $x' \in \mathcal{X}$ to $T_s(x') = y(x, x')$. Then, clearly (28) holds, thereby proving the necessity of the condition for positive capacity in the theorem.

Next, recall that for any AVC, if the capacity is positive, it equals $\min_Q C(W_Q)$ for Q ranging over the probability distributions on the state set $\mathcal{S}$, where W_Q is the Q-mixture of the channels corresponding to the individual states $s \in \mathcal{S}$ (cf. Ahlswede [1]). Clearly, in our case, the channels of form W_Q are the same as the channels compatible with the graph $\mathcal{G}$. Thus the last statement of Theorem 4 follows.

Next, we claim that if the AVC determined by the graph $\mathcal{G}$ has positive capacity, then every strictly positive input distribution P satisfies the Condition DS stated in [13, Definition 3]. This condition states for a general AVC that no distribution Q on $\mathcal{S}$ and channel $U: \mathcal{X} \to \mathcal{S}$ exist such that for every $x' \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$\sum_{x \in \mathcal{X},\, s \in \mathcal{S}} P(x)W(y|x, s)U(s|x') = \sum_{s \in \mathcal{S}} W(y|x', s)Q(s). \quad (29)$$

In our case, $W(y|x, s) = 1$ if $y = T_s(x)$, and 0 otherwise; thus the right-hand side of (29) equals 0 for every x' and y not connected by an edge of $\mathcal{G}$. By the condition for positive capacity that has already been proved, there exist x and $x' \in \mathcal{X}$ such that no $y \in \mathcal{Y}$ connected to x is also connected to x'. Given any channel $U: \mathcal{X} \to \mathcal{S}$, pick $s \in \mathcal{S}$ with $U(s|x') > 0$ and let $y = T_s(x)$. Then the left-hand side of (29) will be positive, establishing our claim.

Finally, by [13, Theorem 3] the result just proved establishes that capacity can be attained by the "typicality decoding rule" defined in [13]. In our case, the latter reduces to the decoder yielding (27), by virtue of the fact that the channels of form W_Q are those compatible with $\mathcal{G}$. This completes the proof of Theorem 4.

Remarks: The fact that the symmetrizability of an AVC implies zero capacity dates back to Blackwell, Breiman and Thomasian [6]; Csiszár and Narayan [12] proved that symmetrizability (though not its deterministic version) is, indeed, equivalent to zero capacity. Condition DS was introduced by Dobrushin and Stambler [15], and the result that it implies achievability of the capacity of the AVC by "typicality" decoding is largely attributable to them. Of course, Theorem 4 could have been alternatively established in a more direct manner, without recourse to general AVC theory. Our preference, instead, for the tools developed in [13] stems from their ability to readily lead to the desired results, thereby demonstrating their power.

V. Open Problems

It is remarked in Section II that if the decoding metrics d and $\tilde{d}$ are "equivalent" in the sense of (2), then $C_d(W) = C_{\tilde{d}}(W)$ for every DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$.

Problem 1: Conversely, if $C_d(W) = C_{\tilde{d}}(W)$ for *all* DMC's $\{W: \mathcal{X} \to \mathcal{Y}\}$, is it true that d and $\tilde{d}$ are "equivalent" *vis a vis* (2)?

Note that the question above, when restricted to a single DMC, has a negative answer.

Problem 2: Given a decoding metric d, does $C_d(W) = C(W)$ imply $C_d^{(1)}(W) = C(W)$?

For a decoding metric such that $d(x, y) = 0$ iff $W(y|x) > 0$, it is easily seen that $C_d^{(1)}(W) = C(W)$ iff (3) holds subject to $P(x) > 0$ for some P which yields $I(P, W) = C(W)$. The sufficiency of this condition for the equality of e.o. capacity and capacity has already been established in Section II. An affirmative answer to Problem 2 with d as above would imply an affirmative answer to the following problem.

Problem 3: Is the sufficient condition above for $C_{eo}(W) = C(W)$ necessary too?

The next two problems ask if certain standard results on the capacity of a DMC extend to d-capacity.

Problem 4: Does the strong converse hold for a DMC? Namely, for codes with d-decoding and of rate approaching some $R > C_d(W)$ as the block length increases, does the probability of error necessarily go to 1?

Problem 5: If $R < C_d(W)$, do there exist codes with d-decoder, rate approaching R, and probability of error decaying to zero exponentially as the block length goes to ∞?

As an immediate extension of Proposition 1, the property required in Problem 5 does hold for $R < C_d^{(\infty)}(W)$. Thus $C_d^{(\infty)}(W) = C_d(W)$ is a sufficient condition for an affirmative answer to Problem 5. This latter condition is known to be satisfied if d is an e.o. metric, by Theorem 3.

Problem 6: Does $C_d^{(\infty)}(W)$ equal $C_d(W)$ for every DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ and decoding metric d?

A natural modification of the d-decoder is the (d, τ)-threshold decoder, which accepts codeword $\boldsymbol{x}^{(i)}$ iff $d(\boldsymbol{x}^{(i)}, \boldsymbol{y}) \leq \tau$ and $d(\boldsymbol{x}^{(j)}, \boldsymbol{y}) > \tau$ for every $j \neq i$. The threshold d-capacity of a DMC can then be defined as the largest rate attainable by codes with (d, τ_n)-threshold decoders for suitable thresholds τ_n. Clearly, threshold d-capacity can never exceed d-capacity. Further, $C_d^{(1)}(W)$, and consequently $C_d^{(\infty)}(W)$, constitute lower bounds on threshold d-capacity too (cf. [10], [22]). Hence an affirmative answer to Problem 6 would imply an affirmative answer to the following problem.

Problem 7: Is the threshold d-capacity of a DMC $\{W: \mathcal{X} \to \mathcal{Y}\}$ always equal to its d-capacity?

We show in Section IV that the e.o. capacity of a special kind of deterministic AVC (determined by a bipartite graph) is equal to the capacity of the AVC (for deterministic codes and the average probability of error criterion). It remains unknown whether the same holds for every deterministic AVC.

Problem 8: It the e.o. capacity of a deterministic AVC always equal to its capacity?

VI. Appendix

Proof of Lemma 1: To show that $P_n \to P, W_n \to W$ implies

$$\lim_n I_d(P_n, W_n) = I_d(P, W) \tag{A1}$$

denote by P^* (resp. P_n^*) the joint distribution of X and Y attaining the minimum in (9) for (P, W) (resp. (P_n, W_n)).

Note first that any sequence of the positive integers contains a subsequence n_k such that $P_{n_k}^* \to \tilde{P}$, say. Clearly, $\tilde{P}$ is the joint distribution of random variables X and Y satisfying the constraints in (9); this proves that

$$\liminf_n I_d(P_n, W_n) \geq I_d(P, W). \tag{A2}$$

To complete the proof of (A1), it suffices to show the existence of distributions $\tilde{P}_n \to P^*$ on $\mathcal{X} \times \mathcal{Y}$ such that $\tilde{P}_n$ is the joint distribution of random variables X and Y satisfying the constraints in (9), with P and W replaced by P_n and W_n. Such $\tilde{P}_n$ can be given by

$$\begin{aligned}\tilde{P}_n(x, y) = {} & P_n(x)W_n(y|x) + (1 - \epsilon_n)(P^*(x, y) \\ & - P(x)W(y|x))\end{aligned} \tag{A3}$$

where the sequence of positive numbers $\epsilon_n \to 0$ is chosen so as to satisfy $P_n(x)W_n(y|x) \geq (1 - \epsilon_n)P(x)W(y|x)$ for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$ (possible because $P_n(x)W_n(y|x) \to P(x)W(y|x)$).

To check the remaining assertion of Lemma 1, note that the minimum of $I(X \wedge Y)$ subject to $P_X = P, P_Y = PW, Ed(X, Y) \leq \Delta$ is a convex function of Δ (by the same standard argument used to show the convexity of the rate distortion function). Hence, this function is continuous for $\Delta > 0$ and is strictly decreasing in the interval where it is nonzero, i.e., for Δ less than the value of $Ed(X, Y)$ for independent X and Y with $P_X = P, P_Y = PW$.

Proof of Lemma 2: For the noiseless channel W_0 and d as in the lemma, we have $\tilde{P}W_0^n = \tilde{P}$ and $\Delta(\tilde{P}, W_0^n) = 0$ for all $\tilde{P}$. Hence, $I_d(\tilde{P}, W_0^n)$ equals the minimum of $I(X^n \wedge Y^n)$ over X^n and Y^n such that $P_{X^n} = P_{Y^n} = \tilde{P}$ and $Ed(X^n, Y^n) = 0$. The latter condition means that (X^n, Y^n) is an edge of $\mathcal{G}^n$ with probability 1.

We have to prove that $\max_{\tilde{P}} I_d(\tilde{P}, W_0^n)$ equals the log of the cardinality of the largest subset of $\mathcal{X}^n$ such that no $\boldsymbol{x}$ and $\boldsymbol{x}'$ in this subset are connected by an edge of $\mathcal{G}^n$. Clearly, it suffices to prove this for $n = 1$; the general case then follows by substituting $\mathcal{G}$ by $\mathcal{G}^n$.

Now, $I_d(P, W_0)$ is the minimum of $I(X \wedge Y)$ over X and Y such that $P_X = P_Y = P$ and $Ed(X, Y) = 0$. The maximum of this $I_d(P, W_0)$ over P is known to equal $\log \alpha(\mathcal{G})$ (cf. [9, p. 87, problem 18(b)]). (Formally, there the constraint $Ed_W(X, Y) < \infty$ appears instead of our $Ed(X, Y) = 0$, where $d_W(x, y) = \infty$ iff our $d_W(x, y)$ is positive; clearly, the two constraints are equivalent.) This completes the proof of the Lemma.

Proof of Lemma 3: Let $X_1, \cdots, X_n$ be independent random variables with $P_{X_i} \in \mathcal{P}$, $i = 1, \cdots, n$, and suppose that m of them are nondegenerate (i.e., $\operatorname{var} X_i > 0$). Without any loss of generality, suppose that these are $X_1, \cdots, X_m$; then $X_i = EX_i$ with probability 1 for $m < i \leq n$. If

$$\sum_{i=1}^n EX_i \geq 0$$

we obtain that

$$\begin{aligned}\Pr\left\{\sum_{i=1}^n X_i \geq 0\right\} &= \Pr\left\{\sum_{i=1}^m (X_i - EX_i) + \sum_{i=1}^n EX_i \geq 0\right\} \\ &\geq \Pr\left\{\sum_{i=1}^m (X_i - EX_i) \geq 0\right\} \geq \frac{1}{4}\end{aligned}$$

if m is sufficiently large, say $m \geq m_0$; the uniform bound in the last step follows, for instance, by Berry's sharpening of the central limit theorem [16, p. 544], relying on the hypothesis that the set of possible distributions of the X_i's is finite. Thus in case $m \geq m_0$, the assertion holds with $\epsilon = 1/4$.

Note further that the set of possible probabilities of the form

$$\Pr\left\{\sum_{i=1}^m X_i \geq a\right\}$$

with $m < m_0$ and arbitrary $a \in \mathbb{R}$ is finite. Let ϵ denote the smallest positive value among these probabilities. Then if (18) holds with this ϵ, and $m < m_0$, we have

$$\Pr\left\{\sum_{i=1}^n X_i \geq 0\right\} = \Pr\left\{\sum_{i=1}^m X_i \geq -\sum_{i=m+1}^n EX_i\right\} = 0$$

which implies (19). This completes the proof of Lemma 3.

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, pp. 10–35, 1980.

[3] R. Ahlswede, N. Cai, and Z. Zhang, "Erasure, list and detection zero-error capacities for low noise and a relation to identification," in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, June–July 1994), p. 380; also, Preprint 93-068, Sonderforschungsbereich 343, *Discrete Strukturen in der Mathematik*, Universität Bielefeld, 1993.

[4] V. B. Balakirsky, "Coding theorems for discrete memoryless channels with given decision rule," in *Lecture Notes in Comp. Sci., Proc. 1st French–Soviet Workshop on Algebraic Coding*, July 1991, pp. 142–150.

[5] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," preprint, 1994.

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

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.
[8] I. Csiszár and J. Körner, "Many coding theorems follow from an elementary combinatorial lemma," in *Proc. 3rd Czechoslovak–Soviet–Hungarian Sem on Information Theory* (Liblice, Czechoslovakia, 1980, pp. 25–44.
[9] ——, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic Press, 1981.
[10] ——, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–12, 1981.
[11] ——, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 57, pp. 87–101, 1981.
[12] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Capacity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, Mar. 1988.
[13] ——, "Capacity and decoding rules for classes or arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 752–769, July 1989.
[14] I. Csiszár, "Channel capacity for minimum distance decoding," in *Open Problems, 4th Joint Swedish–Soviet Workshop on Information Theory*, (Gotland, Sweden, 1989), pp. 8–11.
[15] 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).
[16] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. II. New York: Wiley, 1971.
[17] T. R. M. Fischer, "Some remarks on the role of inaccuracy in Shannon's theory of information transmission," in *Trans. 8th Prague Conf. on Information Theory*, 1971, pp. 211–226.
[18] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.
[19] ——, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.
[20] L. Gargano, J. Körner, and U. Vaccaro, "Sperner capacities," *Graphs and Combinatorics*, vol. 9, pp. 31–46, 1993.
[21] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.
[22] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, MIT, 1983.
[23] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU*, vol. 47, no. 4, pp. 228–239, 1993.
[24] A. Lapidoth, "Information rates for mismatched decoders," in *Proc. 2nd Int. Winter Meeting on Coding and Information Theory* (Essen, Germany, Dec. 1993), pp. 12–15.
[25] ——, "Mismatched decoding for the multiple access channel, in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, June–July 1994), p. 382.
[26] L. Lovàsz, "On the Shannon capacity of a graph," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 1–7, 1979.
[27] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," in *Proc. 1993 IEEE Int. Symp. on Information Theory*, (San Antonio, TX, Jan. 1993), p. 266, also, preprint, 1993.
[28] M. S. Pinsker and A. Sheverdjaev, "Zero error capacity with erasure," *Probl. Peredach. Inform.*, vol. 6, no. 1, pp. 20—24 (in Russian), 1970.
[29] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 8–19, 1956. (Reprinted in *Key Papers in the Development of Information Theory*, D. Slepian, Ed. New York: IEEE Press, 1974.)
[30] M. Simon, J. Omura, R. Sholtz, and B. Levitt, *Spread Spectrum Communications*, vol. I, New York: Computer Science Press, 1985, ch. 4.
[31] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 189–195, Apr. 1966.
[32] I. E. Telatar and R. G. Gallager, "New exponential upper bounds to error and erasure probabilities," in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, June–July 1994), p. 379; also, preprint, 1993.
[33] J. Wolfowitz, *Coding Theorems of Information Theory.* Berlin-Heidelberg, Germany: Springer Verlag, 1961; 2nd ed., 1964; 3rd ed., 1978.
[34] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 453–460, July 1985.