Arbitrarily Varying Channels with States Sequence Known to the Sender

RUDOLF AHLSWEDE

Abstract—**The capacity of the arbitrarily varying channels with states sequence known to the sender is determined. The result is obtained with the help of an elimination technique and a robustification technique. It demonstrates once more the power of these techniques.**

I. Formulation of the Channel Model and the Coding Problem

LET $\mathcal{X}$, $\mathcal{Y}$ be finite alphabets, let $\mathcal{S}$ be a set of states, and let $\mathcal{W} = \{w(\cdot|\cdot|s) : s \in \mathcal{S}\}$ be a set of stochastic $|\mathcal{X}| \times |\mathcal{Y}|$-matrices.

We call the sequence $(\mathcal{W}^n)_{n=1}^{\infty}$ an arbitrarily varying (AV) channel if

$$\mathcal{W}^n = \{w(\cdot|\cdot|s^n) : s^n \in \mathcal{S}^n\} \tag{1.1}$$

$$w(y^n|x^n|s^n) = \prod_{t=1}^{n} w(y_t|x_t|s_t) \tag{1.2}$$

for all

$$x^n = (x_1, \cdots, x_n) \in \mathcal{X}^n,\ y^n \in \mathcal{Y}^n, \quad \text{and} \quad s^n \in \mathcal{S}^n.$$

An AV channel can be viewed as a model for a discrete memoryless transmission system, which depends on a parameter or state, that may change within $\mathcal{S}$ in an arbitrary manner. It also can be viewed as a channel model for a jamming situation in which the jammer chooses the states.

There is a large number of coding problems for these channels because

(a) the sender, receiver, and jammer can have at every time instant t a certain *side information* about the past, present, and future "operation of the system;"
(b) the communicators can often achieve higher capacities with randomized coding strategies; and
(c) the error criteria (maximal versus average) can drastically affect a coding problem.

The reader is advised to consult [2], [7], and [10] for a discussion of these phenomena. A more complete account will be given in [8].

Manuscript received November 23, 1984; revised November 11, 1985. This paper was an invited paper at a Statistical Research Conference dedicated to the memory of Jack Kiefer and Jacob Wolfowitz, held at Cornell University, July 1983, and was presented at the Ninth Prague Conference on Information Theory, Stat. Dec. Fct's, and Random Proc., 1982.

The author is with the Universität Bielefeld, Fakultät für Mathematik, 4800 Bielefeld 1, Federal Republic of Germany.

IEEE Log Number 8608902.

We make here the following assumptions:

(J^-) The jammer has no knowledge of messages or words or letters, which are to be transmitted or have been sent or have been received.
(S^+) The sender knows the *entire sequence* s^n actually used by the jammer before he transmits a message.
(R^-) The receiver has no knowledge of the jammer's (pure or mixed) strategies.

It is convenient to assume also that $\mathcal{S}$ is finite. All our results can be extended by standard approximations (as in [2]) to the case of infinite $\mathcal{S}$. In our analysis we consider cases where the set of strategies available to the jammer for block length n is a subset, say, $\mathcal{P}_n'$, of the set $\mathcal{P}(\mathcal{S}^n)$ of all mixed strategies, that is, of all probability distributions on $\mathcal{S}^n$.

Since our assumptions about the receiver's and sender's knowledge are always the same, a channel model is completely specified by $(\mathcal{P}_n')_{n=1}^{\infty}$. We introduce short names for the channel models of interest to us:

$\mathcal{C}_0$ describes the case where the jammer does not randomize; that is, $\mathcal{P}_n'$ consists of all one-point distributions.
$\mathcal{C}$ describes the case where the jammer can randomize in any manner, that is, $\mathcal{P}_n' = \mathcal{P}(\mathcal{S}^n)$.
$\mathcal{C}_Q$, $Q \subset \mathcal{P}(\mathcal{S})$ describes the case where the jammer can randomize in a special manner, namely, he can use only strategies $q^n = \prod_1^n q$, $n \in \mathbb{N}$, with $q \in Q$.

In particular, if $Q = \{q\}$ has a single element, we write also $\mathcal{C}_q$. This describes a channel that Gelfand–Pinsker [1] called a *channel with random parameters*. Their characterization of its capacity, stated in the following as Theorem 1, serves as one basic tool for our solution of the capacity problem for $\mathcal{C}$, about which they write [1, p. 20]:

> *"This means that one has to construct codes which fit for every set of states* $s_1, \cdots, s_t, \cdots$. *This approach makes the problem much more difficult, and one has no final answer."*

We need a few definitions.

Codes, Errors, and Capacities

An (n, N, λ)-code for a channel model specified by jammer's strategies $(\mathcal{P}_n'^{\infty})_{n=1}$ is a system $\{(u_i(s^n), D_i) : s^n$

0018-9448/86/0900-0621$01.00

$\in \mathcal{S}^n, 1 \le i \le N\}$ with $u_i(s^n) \in \mathcal{X}^n$, $D_i \subset \mathcal{Y}^n$ pairwise disjoint, and

$$\sum_{s^n} w(D_i|u_i(s^n)|s^n)q_n(s^n) \ge 1 - \lambda \quad \text{for} \quad 1 \le i \le N, q_n \in \mathcal{P}_n'. \quad (1.3)$$

If we consider the average probability of error criterion instead of the maximum error criterion, then (1.3) is replaced by

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{s^n} w(D_i|u_i(s^n)|s^n)q_n(s^n) \ge 1 - \bar{\lambda} \quad \text{for} \quad q_n \in \mathcal{P}_n', \quad (1.4)$$

and we speak of an $(n, N, \bar{\lambda})$-code for $(\mathcal{P}_n')_{n=1}^{\infty}$.

The capacities for $\mathcal{C}_0, \mathcal{C}, \mathcal{C}_Q, \mathcal{C}_q$ in the case of maximum (respectively average) probability of errors are denoted by C_0, C, C_Q, C_q (respectively $\bar{C}_0, \bar{C}, \bar{C}_Q, \bar{C}_q$). Since $w(D_i|u_i(s^n)|s^n) \ge 1 - \lambda$ (respectively

$$1/N\textstyle\sum_{i=1}^{N} w(D_i|u_i(s^n)|s^n) \ge 1 - \lambda)$$

for all $s^n \in \mathcal{S}^n$ implies $\sum_{s^n} w(D_i|u_i(s^n)|s^n)q(s^n) \ge 1 - \lambda$ (respectively $1/N\sum_{i=1}^{N}\sum_{s^n} w(D_i|u_i(s^n)|s^n)q(s^n) \ge 1 - \lambda$) we have

$$C_0 = C, \qquad \bar{C}_0 = \bar{C}. \quad (1.5)$$

Furthermore, it follows from

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{s^n} w(D_i|u_i(s^n)|s^n)q(s^n) \ge 1 - \lambda$$

that for a suitable half of the indices i

$$\sum_{s^n} w(D_i|u_i(s^n)|s^n)q(s^n) \ge 1 - 2\lambda$$

and therefore

$$C_q = \bar{C}_q. \quad (1.6)$$

We shall see later that actually $C = \bar{C}$ in contrast to the behavior of AV channels in other cases (see [2]).

The Result for C_q

Let $\mathcal{R}_q$ be the set of all triples (U, S, X) of random variables with values in $\mathcal{U}$, $\mathcal{S}$ and $\mathcal{X}$, respectively, where $\mathcal{U}$ is a given finite set, and with a joint probability distribution P_{USX} such that the marginal distribution of S is q.

To every triple $(U, S, X) \in \mathcal{R}_q$ assign the quadruple (U, S, X, Y) with

$$P_{USXY}(u, s, x, y) = P_{USX}(u, s, x)w(y|x|s).$$

For every finite set $\mathcal{X}$ denote its cardinality by $|\mathcal{X}|$.

The following result of [1] and several extensions thereof were also obtained in [6]. All this work originated with the study of storage in defective memory cells [5].

Theorem 1:

$$C_q = \max_{(U,S,X)\in\mathcal{R}_q} [I(U \wedge Y) - I(U \wedge S)]$$

for a $\mathcal{U}$ with $|\mathcal{U}| \le |\mathcal{X}| + |\mathcal{S}|$.

One of the discoveries of [2] was that the problem of finding the average error capacity for AV channels with states unknown to sender and receiver can be divided into two subproblems, namely, the problem of *positivity of the capacity* and the problem of the *capacity formula*. This dichotomy is present in most capacity problems for AV channels. Moreover, it is a remarkable phenomenon that there is a certain trade-off between the complexity of the two subproblems. By our experience a capacity theorem is very hard to establish if one of the subproblems turns out to be easy. An example for this is the zero-error capacity problem. In our present case none of the subproblems is easy.

The mathematics of the subproblems can be quite different. The positivity part leads to several nice geometrical problems for cartesian products of convex sets, which seem to be of independent interest. This will be the main subject of [8].

The formula part of our problem is a demonstration for the power of our robustification technique [3, part II] and our elimination technique [2].

The reader is advised to study [2], where many of the techniques of the present paper, in particular the elimination technique, were first developed. He/she may also consult the paragraph on arbitrarily varying channels in [10]. Familiarity with [3] is not necessary for an understanding of this paper.

II. The Results

The *robustification technique* exploits the simple fact that the probabilistic structure of channels is invariant under permutations of the components. Once one has a compound channel coding theorem it is merely an exercise to derive via the *robustification technique* in conjunction with the *elimination technique* a capacity formula for the corresponding AV channel if the capacity is positive.

Thus in the present situation we first need an extension of Theorem 1 to the compound case $\mathcal{C}_Q$. Here the sender knows s^n, but not q. However, he can inform the receiver about the type of s^n.

We state the result as follows.

Proposition 1:

$$C_Q = \inf_{q\in Q} C_q, Q \subset \mathcal{P}(\mathcal{S}), \qquad \text{if } C > 0,$$

and rates below C_Q can be achieved with exponentially small error probability in the block length.

Actually, the formula is valid also if $C = 0$, but this will not be used and therefore is not proved here. The positivity part of the problem is settled in the following nonobvious lemma.

Separation Lemma: The condition $C > 0$ is equivalent to the separation condition

(T) all $w \in \mathcal{W}$ have two distinct row vectors.

(This means, of course, that all DMC's with transmission matrix $w \in \mathcal{W}$ have positive capacity.)

Our main result is the following.

Capacity Theorem: The capacity C of the AV channel with state sequence known to the sender before transmission is given by

$$C = \min_{q \in \mathcal{P}(\mathcal{S})} C_q.$$

Moreover, the capacities for maximum and average probabilities of errors are equal, that is, $C = \bar{C}$.

Remarks

In contrast to the capacity theorem of [2], here the formula for the capacity is valid regardless of whether $C > 0$ or $C = 0$.

[7, Theorem 6.2.3] says that the condition

$$(\tilde{T}) \quad \min_{q, q' \in \mathcal{P}(\mathcal{S})} \sum_{y} \left| \sum_{s} w(y|u(s)|s) q(s) - \sum_{s} w(y|v(s)|s) q'(s) \right| > 0$$

for suitable functions $u, v: \mathcal{S} \to \mathcal{X}$,

is sufficient for $C > 0$. However, this condition is not necessary for $C > 0$ to hold, and thus not exact. *Example 1*: For $\mathcal{W} = \left\{ \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} \right\}$ is clearly non-$(\tilde{T})$, but by the separation lemma $C > 0$.

It seems to us that the aforementioned capacity theorem is one of the most complex coding theorems ever proved. Its proof not only involves the techniques of AVC theory but also some of the most advanced techniques from multiuser theory entering via the proof of Theorem 1. Fifteen years ago such a capacity theorem must have been out of reach, and now it serves almost only as a demonstration for the power of certain methods. It is even conceivable that soon a much simpler proof will be found. This shows that there is hope also for several of the harder problems in multiuser theory, which seem to resist all efforts for their solution. Some problems can be solved only at the right time; the time is right if the methods are mature.

Channels with random parameters have been considered already by Shannon [9]. In [7] they are called "channels with stochastically selected states." The kinds of partial side information by the sender and/or receiver studied in [9] and [7] lead to mathematically less sophisticated coding problems.

Thanks are due to a referee for asking whether C can really be smaller than $\min_{w \in \mathcal{W}} C(w)$, where $C(w)$ is the capacity of the DMC with transmission matrix w. We show in the Appendix that in the deterministic case, that is, all $w \in \mathcal{W}$ are 0–1-matrices, $C = \min_{w \in \mathcal{W}} C(w)$, and that already for Example 1 $C < \min_{w \in \mathcal{W}} C(w)$.

III. Proof of the Separation Lemma

Before we enter the formal proof we describe the ideas which led to the characterization of the positivity of C in terms of condition (T). For now, and also later, we use for any finite set $\mathcal{X}$ the following *notation*: $\mathcal{P}(\mathcal{X}) \triangleq$ set of all distributions on $\mathcal{X}$; $\mathcal{P}(n, \mathcal{X}) \triangleq \{ p \in \mathcal{P}(\mathcal{X}) : p(z)n = n_z$ is integral for all $z \in Z \}$; $z^n = (z_1, \cdots, z_n) \in \mathcal{X}^n$ is said to be of type q (or $(q,0)$-typical in the terminology of [3]) if $|\{t : z_t = z\}| = q(z)n$ for all $z \in Z$, and $\mathcal{X}^n(q)$ is the set of sequences in $\mathcal{X}^n$ of type q, $q \in \mathcal{P}(n, \mathcal{X})$.

Proof: Now obviously (T) is necessary for C to be positive; however, sufficiency is by no means obvious and to be proved.

A further observation is the fact that the condition

$$(\tilde{T}_\infty) \quad \min_{q, q' \in \mathcal{P}(\mathcal{S}^n)} \sum_{y^n} \left| \sum_{s^n} w(y^n|u(s^n)|s^n) q(s^n) - \sum_{s^n} w(y^n|v(s^n)|s^n) q'(s^n) \right| > 0$$

for some n and suitable functions $u, v: \mathcal{S}^n \to \mathcal{X}^n$,

which is a natural extension of condition $(\tilde{T})$, is not only necessary, but, as a consequence of [4, Theorem 1] ([7, Theorem 6.2.1]), also sufficient for C to be positive. However, condition $(\tilde{T}_\infty)$ is a "nonsingle letter" condition and therefore hard to verify.

Since knowledge of s^n implies knowledge of the type of s^n, it seems reasonable to consider the (seemingly) stronger condition

(T_1) There exist functions $u, v: \mathcal{P}(\mathcal{S}) \times \mathcal{S} \to \mathcal{X}$ with

$$\min_{q, q' \in \mathcal{P}(\mathcal{S})} \sum_{y} \left| \sum_{s} w(y|u(q, s)|s) q(s) - \sum_{s} w(y|v(q', s)|s) q'(s) \right| > 0.$$

It can be seen, though it will not be used, that (T_1) implies $(\tilde{T}_\infty)$ and a fortiori also (T). However, for Example 3 in the Appendix, which satisfies (T), (T_1) does not hold!

Now we take advantage of the fact that for the present channel randomization in the encoding gives no improvement. Indeed, suppose that $\{(r_i, D_i) : 1 \le i \le N\}$ is a code with maximum probability of error λ, where r_i is the randomized encoding function and D_i the decoding set for message i, then

$$\sum_{x^n} w(D_i|x^n|s^n) r_i(x^n|s^n) \le \max_{x^n} w(D_i|x^n|s^n) \quad (3.1)$$

and we can therefore define, for every s^n, $u_i(s^n)$ as the maximizing x^n. Thus we obtain a code $\{(u_i(s^n), D_i) : s^n \in \mathcal{S}^n, 1 \le i \le N\}$ with a maximum probability of error not exceeding λ.

We can, therefore, look for conditions characterizing positivity of the capacity under randomized encoding. It is natural to weaken condition (T_1) to the condition (T_1^*). There exist sets of random variables $\{U(q, s) : (q, s) \in \mathcal{P}(\mathcal{S}) \times \mathcal{S}\}$ and $\{V(q, s) : (q, s) \in \mathcal{P}(\mathcal{S}) \times \mathcal{S}\}$, each

random variable taking values in $\mathcal{X}$, such that

$$\min_{q,q' \in \mathcal{P}(\mathcal{S})} \sum_y \left| \sum_s \mathbf{E} w(y|U(q,s)|s) q(s) - \sum_s \mathbf{E} w(y|V(q',s)|s) q'(s) \right| > 0.$$

Fortunately, conditions (T) and (T_1^*) are equivalent and (T_1^*) implies $C > 0$. These facts will be stated as Lemmas 1 and 2. With their proofs and the fact that (T) is necessary for C to be positive we have also proved the separation lemma.

Lemma 1: The conditions (T) and (T_1^*) are equivalent. For the proof we need the following elementary lemma.

Selection Lemma: Let J be the set of finite intervals in $\mathbf{R}$ of a length at least c. For any $\epsilon < c/2$ there exist two (selection) functions $f: J \to \mathbf{R}$, $g: J \to \mathbf{R}$ with the properties

$$f(I) \in I, \; g(I) \in I \quad \text{for all} \quad I \in J \tag{3.2}$$

$$|f(I) - g(I')| \geq \epsilon \quad \text{for all} \quad I, I' \in J. \tag{3.3}$$

(There is an obvious generalization to $\mathbf{R}^m$.)

Proof: Define $A = \{2m\epsilon : m \text{ integral}\}$ and $B = \{(2m+1)\epsilon : m \text{ integral}\}$. Since $2\epsilon < c$, for all $I \in J$ we have $A \cap I \neq \emptyset$ and $B \cap I \neq 0$. Define $f(I) = \min\{a : a \in A \cap I\}$ and $g(I) = \min\{b : b \in B \cap I\}$ for all $I \in J$. Thus (3.2) holds. Since for $a \in A$, $b \in B$ necessarily $|a - b| \geq \epsilon$, also (3.3) holds.

Proof of Lemma 1: Obviously non-(T) implies non-(T_1^*), because there is a matrix $w(\cdot|\cdot|s_0)$ with identical rows and for $q = q'$ assigning probability 1 to s_0 the minimization gives the value 0, regardless of the choice of the random variables.

(T) implies (T_1^*): By (T) every matrix $w(\cdot|\cdot|s) \in \mathcal{W}$ has two distinct row vectors with indices a_s and b_s. We construct now a line segment in $\mathbf{R}^{|\mathcal{Y}|}$ for each $q \in \mathcal{P}(\mathcal{S})$. For an s^* with $q(s^*) = \max_s q(s)$ let $I(q)$ be the line segment joining the points

$$\sum_{s \in S} w(\cdot|a_s|s) q(s)$$

and

$$\sum_{s \neq s^*} w(\cdot|a_s|s) q(s) + w(\cdot|b_{s^*}|s^*) q(s^*)$$

in $\mathbf{R}^{|\mathcal{Y}|}$. The length of each $I(q)$ is greater than (or equal to)

$$\min_{s^*} \left(\sum_y |w(y|a_{s^*}|s^*) - w(y|b_{s^*}|s^*)|^2 \right)^{1/2} |\mathcal{S}|^{-1} > 0.$$

Since there are only $|\mathcal{S}|$ many pairs (a_s, b_s), all the line segments $I(q)$, $q \in \mathcal{P}(\mathcal{S})$, take one out of $|\mathcal{S}|$ many directions. Therefore we can find a line $\mathcal{L}$ in $\mathbf{R}^{|\mathcal{S}|}$ such that the projections $\text{proj}_{\mathcal{L}} I(p)$ of the line segments $I(q)$ on $\mathcal{L}$ all have a length greater than a constant $c > 0$. By the selection lemma there are functions f, g with

$$\|f(\text{proj}_{\mathcal{L}} I(q)) - g(\text{proj}_{\mathcal{L}} I(q'))\|_2 \geq c/3 \quad \text{for all} \quad q, q' \in \mathcal{P}(\mathcal{S}).$$

Hence there are also selection functions F, G for the original line segments of the form

$$F(I(q)) = \sum_{s \neq s^*} w(\cdot|a_s|s) q(s) + \left[\alpha_q w(\cdot|a_{s^*}|s^*) + (1 - \alpha_q) w(\cdot|b_{s^*}|s^*) \right] q(s^*)$$

$$G(I(q)) = \sum_{s \neq s^*} w(\cdot|a_s|s) q(s) + \left[\beta_q w(\cdot|a_{s^*}|s^*) + (1 - \beta_q) w(\cdot|b_{s^*}|s) \right] q(s^*)$$

with

$$\|F(I(q)) - G(I(q'))\|_2 \geq c/3 \quad \text{for all} \quad q, q' \in \mathcal{P}(\mathcal{S}).$$

(T_1^*) follows now with the choices

$$U(q,s) = \begin{cases} a_s & \text{for } s \neq s^* \\ \left.\begin{array}{ll} a_{s^*} & \text{with probability } \alpha_q \\ b_{s^*} & \text{with probability } 1 - \alpha_q \end{array}\right\} & \text{for } s = s^* \end{cases}$$

$$V(q,s) = \begin{cases} a_s & \text{for } s \neq s^* \\ \left.\begin{array}{ll} a_{s^*} & \text{with probability } \beta_q \\ b_{s^*} & \text{with probability } 1 - \beta_q \end{array}\right\} & \text{for } s = s^* \end{cases}$$

Lemma 2: (T_1^*) implies $C > 0$.

Proof: We have shown in (3.1) that we can randomize in the encoding. Again by [4, Theorem 1] it suffices to show that two messages can be transmitted with small $(< 1/2)$ error probabilities.

Now the sender, having observed s^n of type q, encodes message 1 as

$$(U(q,s_1), \cdots, U(q,s_n))$$

and message 2 as

$$(V(q,s_1), \cdots, V(q,s_n)).$$

The receiver receives Y_1^n respectively Y_2^n. By the law of large numbers, with a probability tending to one as n tends to infinity the type of Y_1^n approaches the set $\{\mathbf{E} w(\cdot|U(q,s)|s) q(s) : q \in \mathcal{P}(\mathcal{S})\}$ and the type of Y_2^n approaches the set $\{\mathbf{E} w(\cdot|V(q,s)|s) q(s) : q \in \mathcal{P}(\mathcal{S})\}$. By (T_1^*), for sufficiently large n, with high probability, those types are a guaranteed distance apart and therefore, with high probability, the reciever decodes those messages correctly.

IV. Proof of the Capacity Theorem

A. Proof of Proposition 1

By Theorem 1 and a standard concatenation argument (or also by inspection of the proof of Theorem 1) every rate $R < C_Q$ can be achieved for all large m uniformly in $q \in Q$ with an exponentially small error probability, that is, there exist an $\epsilon_1(R)$ and codes $\{(u_i^q, D_i^q) : 1 \leq i \leq M;\ M \geq e^{Rm}\}$ with

$$\sum_{s^m \in \mathcal{S}^m} w\left((D_i^q)^c | u_i^q(s^m) | s^m \right) q^m(s^m) \leq e^{-\epsilon_1 m}$$

for $1 \leq i \leq M$, $q \in Q$, and all large m. (4.1)

However, in our compound coding problem sender and receiver do not know q. We can overcome this difficulty as follows. Set $n \triangleq (1+\delta)m$. Observing s^m, the sender knows its type p. Using δm digits, for instance, after the block of length m, the sender can inform (since $C > 0$) the decoder about p. Since $|\mathcal{P}(m, \mathcal{S})| \le (m+1)^{|\mathcal{S}|}$, this can be done with a code $\{(V^P, B^P): p \in \mathcal{P}(m, \mathcal{S})\}$ of negligible rate, that is, with a very small δ and with an error probability less than $e^{-\epsilon_2(\delta)m}$, where $\epsilon_2(\delta) > 0$.

Based on the previous codes, we define now a code *not depending on* $q \in Q$ for blocklength n. For this we assign first to every $p \in \mathcal{P}(m, \mathcal{S})$ and $p' \in Q$ such that

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{s^m \in \mathcal{S}^m(p)} w\left(\left(D_i^{p'}\right)^c | u_i^{p'}(s^m) | s^m\right) \le \inf_{q \in Q} \frac{1}{M}\sum_{i=1}^{M}\sum_{s^m \in \mathcal{S}^m(p)} w\left(\left(D_i^{q}\right)^c | u_i^{q}(s^m) | s^m\right) + e^{-\epsilon_3 m}, \quad \epsilon_3 > 0. \tag{4.2}$$

Now for every $s^n = s^m s_{m+1} \cdots s_n$, $s^m \in \mathcal{S}^m(p)$ we define

$$u_i(s^n) \triangleq u_i^{p'}(s^m) v^p,\ D_i(s^n) = D_i^{p'} \times B^p;\ 1 \le i \le M. \tag{4.3}$$

The receiver uses this decoding system by first deciding upon the type p of s^m and then, with the knowledge of p, he decides on the message i.

We estimate now the average error probability for this code. For any $q \in Q$ by (4.3)

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\sum_{s^n \in \mathcal{S}^n} w\left(D_i(s^n)^c | u_i(s^n) | s^n\right) q^n(s^n) \\
&\le \frac{1}{M}\sum_{i=1}^{M}\sum_{p \in \mathcal{P}(m,\mathcal{S})}\sum_{s^m \in \mathcal{S}^m(p)} w\left(\left(D_i^{p'}\right)^c | u_i^{p'}(s^m) | s^m\right) \cdot q^m(s^m) + e^{-\epsilon_2(\delta)m} \\
&= \frac{1}{M}\sum_{i=1}^{M}\sum_{p \in \mathcal{P}(m,\mathcal{S})}\left(\sum_{s^m \in \mathcal{S}^m(p)} w\left(\left(D_i^{p'}\right)^c | u_i^{p'}(s^m) | s^m\right)\right) \cdot q^m(s^m) + e^{-\epsilon_2(\delta)m}
\end{aligned}$$

(since all sequences in $\mathcal{S}^m(p)$ have the same q^m-probability)

$$\begin{aligned}
&\le \frac{1}{M}\sum_{i=1}^{M}\sum_{p \in \mathcal{P}(m,\mathcal{S})}\left(\sum_{s^m \in \mathcal{S}^m(p)} w\left(\left(D_i^{q}\right)^c | u_i^{q}(s^m) | s^m\right)\right) \cdot q^m(s^m) + e^{-\epsilon_2(\delta)m} + e^{-\epsilon_3 m} \quad \text{(by (4.2))} \\
&\le e^{-\epsilon_1 m} + e^{-\epsilon_2(\delta)m} + e^{-\epsilon_3 m} \quad \text{(by (4.1))}
\end{aligned}$$

B. The Robustification Technique

We use the symmetric group Σ_n, that is, the group of permutations acting on $\{1, 2, \cdots, n\}$. Every $\sigma \in \Sigma_n$ induces a bijection $\pi: \mathcal{S}^n \to \mathcal{S}^n$ defined by $\pi s^n \triangleq (s_{\sigma(1)}, \cdots, s_{\sigma(n)})$ for $s^n = (s_1, \cdots, s_n) \in \mathcal{S}^n$. Π_n denotes the group of these bijections. Their restrictions to $\mathcal{S}^n(p)$, $p \in \mathcal{P}(n, \mathcal{S})$, are also bijective. In [3, part II, ch 5] a robustification technique was formulated as Theorem 6. An analysis of its proof led to the more general theorem that follows.

Theorem Robustification Technique (RT): If $g: \mathcal{S}^n \to [0,1]$ satisfies for an $\alpha \in (0,1)$ the inequality

$$\sum_{s^n \in \mathcal{S}^n} g(s^n) p^n(s^n) > 1 - \alpha, \quad \text{for all } p^n = \prod_1^n p \text{ with } p \in \mathcal{P}(n, \mathcal{S}),$$

then it also satisfies the inequality

$$\frac{1}{n!}\sum_{\pi \in \Pi_n} g(\pi s^n) < 1 - \alpha_n, \quad \text{for all } s^n \in \mathcal{S}^n,$$

where $\alpha_n \triangleq \alpha(n+1)^{|\mathcal{S}|}$.

Proof: The first inequality is equivalent to

$$\sum_{s^n}(1 - g(\pi s^n)) p^n(\pi s^n) < \alpha \quad \text{for } \pi \in \Pi_n,\ p \in \mathcal{P}(n, \mathcal{S}),$$

because π is bijective. Since $p^n(\pi s^n) = p^n(s^n)$, it follows that

$$\sum_{s^n}\left(1 - \frac{1}{n!}\sum_{\pi} g(\pi s^n)\right) p^n(s^n) < \alpha \quad \text{for } p \in \mathcal{P}(n, \mathcal{S}).$$

Here $1 - 1/n!\Sigma_\pi g(\pi s^n) \ge 0$, and therefore the left side is decreased when summing for $s^n \in \mathcal{S}^n(p)$ only. For these s^n $\Sigma_\pi g(\pi s^n)$ is constant and we get

$$\left(1 - \frac{1}{n!}\sum_{\pi} g(\pi s^n)\right) p^n(\mathcal{S}^n(p)) < \alpha \tag{4.4}$$

for every $s^n \in \mathcal{S}^n(p)$ and every $p \in \mathcal{P}(n, \mathcal{S})$.

It is well-known (c.f. [10, p. 30]) that

$$p^n(\mathcal{S}^n(p)) \ge (n+1)^{-|\mathcal{S}|} \quad \text{for } p \in \mathcal{P}(n, \mathcal{S}). \tag{4.5}$$

Now (4.4) and (4.5) imply $(1 - 1/n!\Sigma_\pi g(\pi s^n))(n+1)^{-|\mathcal{S}|} < \alpha$ for all $s^n \in \mathcal{S}^n$, and thus the second aforementioned inequality.

C. Application of Theorem RT

We need the following concepts. A correlated (n, N)-code is specified by a finite probability space (Γ, μ) and a collection $\{(u_i^\gamma(s^n), D_i^\gamma): s^n \in \mathcal{S}^n, 1 \le i \le N\}_{\gamma \in \Gamma}$ of (n, N)-codes. In using such a code, the index γ is chosen according to the random experiment (Γ, μ), and then sender and reciever use the code indexed by γ. Since γ has to be made known to both of them, there must be a common knowledge or correlation in the system. It serves here only as a mathematical tool. The average error is measured by

$$\max_{s^n \in \mathcal{S}^n}\sum_{\gamma \in \Gamma}\mu(\gamma)\frac{1}{N}\sum_{i=1}^{N} w\left(\left(D_i^\gamma\right)^c | u_i^\gamma(s^n) | s^n\right).$$

$\overline{C}_{\text{corr}}$ is the capacity of $\mathcal{C}$ for correlated codes.

Proposition 2:

$$\overline{C}_{\text{corr}} = \min_{q \in \mathcal{P}(\mathcal{S})} C_q.$$

Moreover, rates below $\bar{C}_{\text{corr}}$ can be achieved with exponentially small error probability in the block length.

Proof: By the separation lemma, if $C = 0$, then non-(T) and thus both $\bar{C}_{\text{corr}}$ and $\min_{q \in \mathscr{P}(\mathscr{S})} C_q$ equal 0. In case $C > 0$ by Proposition 1 for $Q \triangleq \mathscr{P}(\mathscr{S})$ there are codes

$$\{(u_i(s^n), D_i) : 1 \le i \le N; s^n \in \mathscr{S}^n\}$$

of a rate arbitrarily close to $\min_{q \in \mathscr{P}(\mathscr{S})} C_q$ with

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{s^n \in \mathscr{S}^n} w(D_i | u_i(s^n) | s^n) q^n(s^n) > 1 - \alpha \quad (4.6)$$

for all $q \in \mathscr{P}(\mathscr{S})$, where α is exponentially small.

For g defined by

$$g(s^n) = \frac{1}{N} \sum_{i=1}^{N} w(D_i | u_i(s^n) | s^n), s^n \in \mathscr{S}^n, \quad (4.7)$$

the inequality (4.6) ensures the validity of the hypothesis of Theorem RT and therefore

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} \frac{1}{N} \sum_{i=1}^{N} w(D_i | u_i(\pi s^n) | \pi s^n) > 1 - \alpha_n. \quad (4.8)$$

By (1.1) and (1.2) this is equivalent to

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} \frac{1}{N} \sum_{i=1}^{N} w(\pi^{-1} D_i | \pi^{-1} u_i(\pi s^n) | s^n) > 1 - \alpha_n. \quad (4.9)$$

But this says that the *correlated* code specified by the collection of codes

$$\{(\pi^{-1} u_i(\pi s^n), \pi^{-1} D_i) : 1 \le i \le N; s^n \in \mathscr{S}^n\}_{\pi \in \Pi_n},$$

and the uniform distribution on Π_n has an average error probability less than α_n. Clearly, $\alpha_n \le e^{-\epsilon n}$ for $\alpha = (n+1)^{-|\mathscr{S}|} e^{-\epsilon n}$.

D. Application of the Elimination Technique (ET)

Theorem ET: If for $\epsilon > 0$, $\lambda \in (0,1)$ and n sufficiently large the *average* probability of error of the correlated code

$$(\Gamma, \mu; \{(u_i^\gamma(s^n), D_i^\gamma) : 1 \le i \le N; s^n \in \mathscr{S}^n\}_{\gamma \in \Gamma})$$

satisfies

$$\max_{s^n} \sum_{\gamma \in \Gamma} \mu(\gamma) \frac{1}{N} \sum_{i=1}^{N} w((D_i^\gamma)^c | u_i^\gamma(s^n) | s^n) \le e^{-\epsilon n}, \quad (4.10)$$

then there exists a $\Gamma^* \subset \Gamma$ and a μ^* such that the correlated code

$$(\Gamma^*, \mu^*, \{(u_i^\gamma(s^n), D_i^\gamma) : 1 \le i \le N; s^n \in \mathscr{S}^n\}_{\gamma \in \Gamma^*})$$

satisfies for its *maximum* probability of error and for Γ^*

$$\max_{1 \le i \le N} \max_{s^n} \sum_{\gamma \in \Gamma^*} \mu^*(\gamma) w((D_i^\gamma)^c | u_i^\gamma(s^n) | s^n) \le \lambda,$$

$$|\Gamma^*| \le n^2. \quad (4.11)$$

Further, if $C > 0$, then (4.11) implies that for blocklength $k = n + \lceil 4/C \log n \rceil$ there exists a code

$$\{(r_i(\cdot | s^k), B_i) : s^k \in \mathscr{S}^k; 1 \le i \le N\} \quad (4.12)$$

with randomized encoding and maximum probability of error less than 2λ.

Proof: This result is proved in [2, sect. 5] in the case where the u_i^γ do not depend on s^n. Since the random variables S_{ji} are defined there for *fixed* s^n, the previous proof for (4.11) *literally* applies, if u_i^γ is replaced by $u_i^\gamma(s^n)$. For the same reason, if q_i in [2, (5.4)] (which corresponds to our r_i) is made dependend on s^k, the previous proof for (4.12) applies. The reader may also consult [10, Lemma 6.8 and Theorem 6.13] and check that the modifications mentioned are of no consequence for the proofs of (4.11) and (4.12).

E. End of proof for the Capacity Theorem

In Theorem ET (4.10) holds for rates arbitrarily close to $\bar{C}_{\text{corr}}$. If $C > 0$, then (4.12) of Theorem ET implies that rates arbitrarily close to $\bar{C}_{\text{corr}}$ can be achieved under randomized encoding with small maximum error probability. By (3.1) the same can be achieved without randomization in the encoding. Therefore $C = \bar{C}_{\text{corr}}$, if $C > 0$.

We know already that $C = 0$ or equivalently non-(T) (separation lemma) implies $\bar{C}_{\text{corr}} = 0$ and thus always

$$C = \bar{C}_{\text{corr}}. \quad (4.13)$$

Since $C \le \bar{C} \le \bar{C}_{\text{corr}}$, we have established the equality of C and $\bar{C}$.

Finally, by Proposition 2 and (4.13) $C = \min_{q \in \mathscr{P}(\mathscr{S})} C_q$.

Appendix

Here we are concerned with calculations and comparisons of channel capacities.

A. The Deterministic Case

Let $\mathscr{W}$ contain 0–1-matrices only. It is instructive to consider the following two examples.

Example 2:

$$\mathscr{X} = \{1,2,3\}, \mathscr{Y} = \{1,2\}, \mathscr{W} = \left\{ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \right\}.$$

Since the sender knows the states, that is, the matrices, at every time instant he can always choose to produce output letter 1 or 2. Therefore $C = \min_w C(w) = 1$.

Example 3:

$$\mathscr{X} = \{1,2\}, \mathscr{Y} = \{1,2,3\}, \mathscr{W} = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\}.$$

Here a simple strategy as before does not exist. In trying to find the capacity, the reader will realize that to solve this seemingly simple problem is already a formidable task. He will then perhaps appreciate the following consequence of the capacity theorem.

Corollary: If $\mathscr{W}$ contains 0–1-matrices only and $d(w)$ is the number of distinct rows of w, then

$$C = \min_w \log d(w).$$

Proof: Let us define $d \triangleq \min_{w \in \mathscr{W}} d(w)$. Since $C(w) = \log d(w)$, we have also $C \le \min_w \log d(w) = \log d$. By the capacity theorem it suffices now to show that $\min_{q \in \mathscr{P}(\mathscr{S})} C_q \ge \log d$. This will be done if we can find for every $q \in \mathscr{P}(\mathscr{S})$ a

triple $(U, S, X) \in \mathcal{R}_q$ with

$$I(U \wedge Y) - I(U \wedge S) \geq \log d.$$

We consider triples with deterministic $P_{X|SU}$, that is, functions $x: \mathcal{S} \times \mathcal{U} \to \mathcal{X}$. Since all $w(\cdot|\cdot|s)$ are deterministic, also the output variable y is deterministic.

We choose now $\mathcal{U} \triangleq \mathcal{Y}$ and shall make our definitions such that $U = Y$. First of all we define for $s \in \mathcal{S}$

$$\mathcal{Y}_s = \{ y: w(y|x|s) = 1, \quad \text{for some} \quad x \in \mathcal{X} \} \subset \mathcal{Y}.$$

Clearly, $|\mathcal{Y}_s| = d(w(\cdot|\cdot|s))$. Next we define for $s \in \mathcal{S}$

$$P_{U|S}(u|s) = \begin{cases} |\mathcal{Y}_s|^{-1} & \text{for } u \in \mathcal{Y}_s \\ 0 & \text{otherwise} \end{cases}$$

and then $x(u, s)$ such that $w(u|x(u,s)|s) = 1$ for $u \in \mathcal{Y}_s$ and arbitrary otherwise, because there $P_{U|S}(u|s) = 0$. These definitions enforce $U = Y$ with probability one. Here $P_U(u) = \sum_s P_{U|S}(u|s) q(s)$.

Now we have

$$\begin{aligned} &I(U \wedge Y) - I(U \wedge S) \\ &\quad = H(U) - H(U) + H(U|S) = \sum_s q(s) \log |\mathcal{Y}_s| \\ &\quad \geq \min_s \log |\mathcal{Y}_s| = \log d. \end{aligned}$$

Shannon found in [9] the capacity of a channel with random parameters in case the sender knows at time t the past and present state.

It is the capacity of a DMC with input alphabet $\mathcal{X}^{|\mathcal{S}|}$, output alphabet $\mathcal{Y}$ and transmission matrix $\tilde{w}$ defined by

$$\tilde{w}(y|x_1, \cdots, x_{|\mathcal{S}|}) = \sum_s q(s) w(y|x_s|s).$$

The row vectors of this matrix for Example 3 with $q = (1/3, 1/3, 1/3)$ are all permutations of $(2/3, 1/3, 0)$ and $(1/3, 1/3, 1/3)$. Since $(1/3, 1/3, 1/3) = 1/2(2/3, 1/3, 0) + 1/2(0, 1/3, 2/3)$ only the six other row vectors are to be considered. By an elementary calculation, $C_{Sh} = \log 3 - h(1/3) = 2/3$, which is smaller than $C = 1$.

We also draw attention to the fact that in Example 3 for the compound channel with transmission matrices $\mathcal{W}$ the capacity, say, C_{comp}, equals 1. However, in Example 2 $C_{comp} = h(1/3)$, whereas $C = C_{Sh}$ (for q as above) $= 1$.

B. *In General* $C < \min_{w \in \mathcal{W}} C(w)$

The corollary tells us that the desired inequality can occur only in the nondeterministic case. The simplest situation we could think of is Example 1 in Section II, where $\mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0,1\}$ and $w(\cdot|\cdot|0) = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}$, $w(\cdot|\cdot|1) = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$. One readily calculates that here for $q = (1/2, 1/2)$

$$C_{Sh} = 1 - h(1/4) \sim 0.19. \tag{A.1}$$

More importantly, we need to know for our comparison the value of $\min_{w \in \mathcal{W}} C(w)$, which by symmetry equals $C\begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}$. Here for the input distribution $(1 - \alpha, \alpha)$ the mutual information $I(\alpha)$ equals $h(\alpha/2) - \alpha$. Since $I'(\alpha) = 1/2 \log(2/\alpha - 1) - 1 = 0$ has the solution $\alpha = 2/5$ and $I''(2/5) < 0$, we get

$$C\begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix} = h(1/5) - 2/5 \sim 0.32. \tag{A.2}$$

Since our capacity C is invariant under permutation of rows in matrices, we could (equivalently) consider the channel $\mathcal{W}' = \left\{ \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}, \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} \right\}$. For curiosity we mention that by symmetry its compound capacity $C_{comp}(\mathcal{W}') = I(1/2) = h(1/4) - 1/2$ and thus

$$C_{comp}(\mathcal{W}') \sim 0.31. \tag{A.3}$$

We show now that for Example 1

$$C \leq 0.28. \tag{A.4}$$

We obtain this bound by proving that for $q = (1/2, 1/2)$ $C_q \leq 0.28$. For this we use the following proposition.

[1, Proposition 1]: For $(U, S, X) \in \mathcal{R}_q$ the function $I(U \wedge Y) - I(U \wedge S)$ is a) for fixed $P_{X|US}$ $\cap$-convex in $P_{U|S}$, and b) for fixed $P_{U|S}$ $\cup$-convex in $P_{X|US}$. We know already from Theorem 1 that in the formula for C_q it suffices to consider U with values in $\mathcal{U}$ such that c) $|\mathcal{U}| \leq |\mathcal{X}| + |\mathcal{S}|$.

We add two simple observations, which are useful in calculating C_q.

Observation 1: The labelling of the elements in $\mathcal{U}$ is irrelevant.

Observation 2: Suppose that for a fixed $P_{X|US}$ we have

$$P_{X|US}(x|u_1 s) = P_{X|US}(x|u_2 s), \quad \text{for all} \quad x \in \mathcal{X}, s \in \mathcal{S}. \tag{A.5}$$

Then we can identify u_1 and u_2 and still achieve the same rate.

Proof: Suppose that also $P_{U|S}$ and $P_s = q$ are specified. Then we can define

$$P_{U'|S}(u|s) = \begin{cases} P_{U|S}(u|s) & \text{for } u \neq u_1, u_2 \\ P_{U|S}(u_2|s) & \text{for } u = u_1 \\ P_{U|S}(u_1|s) & \text{for } u = u_2 \end{cases} \tag{A.6}$$

and $$P_{X|US} = P_{X|U'S}.$$

Hence by (A.5) and Observation 1

$$I(U' \wedge Y) - I(U' \wedge S) = I(U \wedge Y) - I(U \wedge S). \tag{A.7}$$

Define now

$$P_{\tilde{U}|S} = \tfrac{1}{2} P_{U|S} + \tfrac{1}{2} P_{U'|S} \quad \text{and}$$
$$P_{X|\tilde{U}S} = P_{X|US}. \tag{A.8}$$

Then a) of Proposition 1 and (7) imply

$$I(\tilde{U} \wedge \tilde{Y}) - I(\tilde{U} \wedge S) \geq I(U \wedge Y) - I(U \wedge S). \tag{A.9}$$

Finally define

$$P_{U^*|S}(u|s) = \begin{cases} P_{\tilde{U}|S}(u|s) & \text{for } u \neq u_1, u_2 \\ P_{\tilde{U}|S}(u_1|s) + P_{\tilde{U}|S}(u_2|s) & \text{for } u = u^* \text{ (a new element)} \end{cases}$$

and

$$P_{X|U^*S} = P_{X|US}. \tag{A.10}$$

By (5) $I(U^* \wedge Y^*) = I(\tilde{U} \wedge \tilde{Y})$ and by the data processing lemma $I(U^* \wedge S) \leq I(\tilde{U} \wedge S)$. These relations and (9) give

$$I(U^* \wedge Y) - I(U^* \wedge S) \geq I(U \wedge Y) - I(U \wedge S). \tag{A.11}$$

It was explained already in [1] that by b) it suffices to consider deterministic $P_{X|US}$, that is, functions $x: \mathcal{S} \times \mathcal{U} \to \mathcal{X}$. Since in our case $\mathcal{S} = \{0,1\}$, $\mathcal{X} = \{0,1\}$ and by c) $\mathcal{U} = \{1,2,3,4\}$, there are 2^8 such functions. However, by Observations 1 and 2 these possibilities can be reduced to one, namely, to

$$x(u, s): \begin{array}{c|cccc} s \backslash u & 1 & 2 & 3 & 4 \\ \hline 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{array}. \tag{A.12}$$

In its analysis we use the abbreviations

$$A = P_{U|S}, \qquad B = P_{Y|U}. \tag{A.13}$$

A is generally of the form

$$A = \begin{pmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 \\ \delta_1 & \delta_2 & \delta_3 & \delta_4 \end{pmatrix}. \tag{A.14}$$

Recall that $q = P_S = (1/2, 1/2)$, and thus by (12)–(14)

$$B = \begin{pmatrix} \dfrac{\epsilon_1}{2(\epsilon_1+\delta_1)} & 1 - \dfrac{\epsilon_1}{2(\epsilon_1+\delta_1)} \\ 1 - \dfrac{\delta_2}{2(\epsilon_2+\delta_2)} & \dfrac{\delta_2}{2(\epsilon_2+\delta_2} \\ \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{\epsilon_4}{\epsilon_4+\delta_4} & \dfrac{\delta_4}{\epsilon_4+\delta_4} \end{pmatrix}. \tag{A.15}$$

Define now

$$A' = \begin{pmatrix} \delta_2 & \delta_1 & \delta_3 & \delta_4 \\ \epsilon_2 & \epsilon_1 & \epsilon_3 & \epsilon_4 \end{pmatrix}. \tag{A.16}$$

Notice that in A we have exchanged ϵ_1 with δ_2, ϵ_2 with δ_1, ϵ_3 with δ_3, and ϵ_4 with δ_4. Therefore we get

$$B' = \begin{pmatrix} \dfrac{\delta_2}{2(\epsilon_2+\delta_2)} & 1 - \dfrac{\delta_2}{2(\epsilon_2+\delta_2)} \\ 1 - \dfrac{\epsilon_1}{2(\epsilon_1+\delta_1)} & \dfrac{\epsilon_1}{2(\epsilon_1+\delta_1)} \\ \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{\delta_4}{\epsilon_4+\delta_4} & \dfrac{\epsilon_4}{\epsilon_4+\delta_4} \end{pmatrix}. \tag{A.17}$$

Since $P_U(i) = (\epsilon_i + \delta_i)/2$, $1 \le i \le 4$, and since

$$P_{U'}(1) = \frac{\epsilon_2+\delta_2}{2}, \quad P_{U'}(2) = \frac{\epsilon_1+\delta_1}{2},$$

$$P_{U'}(i) = \frac{\epsilon_i+\delta_i}{2} \qquad (i = 3,4), \tag{A.18}$$

inspection of B and B' shows that $I(U \wedge Y) = I(U' \wedge Y')$.

Clearly, also $I(U \wedge S) = I(U' \wedge S)$, and therefore for $\tilde{A} = 1/2A + 1/2A'$ by a)

$$I(\tilde{U} \wedge \tilde{Y}) - I(\tilde{U} \wedge S) \ge I(U \wedge Y) - I(U \wedge S). \tag{A.19}$$

The quantity to the left has symmetry properties which make it possible to calculate

$$C_q = \max_{\epsilon_i, \delta_i} I(\tilde{U} \wedge \tilde{Y}) - I(\tilde{U} \wedge S)$$

if one is guided by the idea that ϵ_3, ϵ_4, δ_3 and δ_4 should be zero. Now by definition of $\tilde{A}$

$$\tilde{A} = \begin{pmatrix} \dfrac{\epsilon_1+\delta_2}{2} & \dfrac{\epsilon_2+\delta_1}{2} & \dfrac{\epsilon_3+\delta_3}{2} & \dfrac{\epsilon_4+\delta_4}{2} \\ \dfrac{\delta_1+\epsilon_2}{2} & \dfrac{\delta_2+\epsilon_1}{2} & \dfrac{\epsilon_3+\delta_3}{2} & \dfrac{\epsilon_4+\delta_4}{2} \end{pmatrix} \tag{A.20}$$

and by the same calculation as the one which led from A to B, we obtain now

$$\tilde{B} = \begin{pmatrix} \dfrac{\gamma}{2\eta} & 1 - \dfrac{\gamma}{2\eta} \\ 1 - \dfrac{\gamma}{2\eta} & \dfrac{\gamma}{2\eta} \\ \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{1}{2} & \dfrac{1}{2} \end{pmatrix},$$

where $\eta \triangleq \dfrac{\epsilon_1+\delta_1+\epsilon_2+\delta_2}{2}$ and $\gamma \triangleq \dfrac{\epsilon_1+\delta_2}{2}$. (A.21)

Furthermore, since $q = (1/2, 1/2)$, from (20)

$$P_{\tilde{U}} = \left(\eta/2, \eta/2, \frac{\epsilon_3+\delta_3}{2}, \frac{\epsilon_4+\delta_4}{2}\right). \tag{A.22}$$

This and (21) imply

$$P_{\tilde{Y}} = \left(\frac{1}{2}, \frac{1}{2}\right). \tag{A.23}$$

Now

$$H(\tilde{Y}) - H(\tilde{Y}|\tilde{U}) = 1 - \left[\eta h\left(\frac{\gamma}{2\eta}\right) + 1 - \eta\right]$$

and

$$\begin{aligned} H(\tilde{U}) - H(\tilde{U}|S) &= H\left(\eta/2, \eta/2, \frac{\epsilon_3+\delta_3}{2}, \frac{\epsilon_4+\delta_4}{2}\right) \\ &\quad - H\left(\gamma, \eta-\gamma, \frac{\epsilon_3+\delta_2}{2}, \frac{\epsilon_4+\delta_4}{2}\right) \\ &= -\eta \log \eta/2 + \gamma \log \gamma \\ &\quad + (\eta-\gamma)\log(\eta-\gamma) \end{aligned}$$

and thus

$$\begin{aligned} I(\tilde{U} \wedge \tilde{Y}) - I(\tilde{U} \wedge S) &= \eta\left(1 - h\left(\frac{\gamma}{2\eta}\right)\right) + \eta \log \eta/2 \\ &\quad - \gamma \log \gamma - (\eta-\gamma)\log(\eta-\gamma) \\ &\triangleq T(\gamma, \eta), \end{aligned} \tag{A.24}$$

say, and

$$C_q = \max_{0 \le \gamma \le \eta \le 1} T(\gamma, \eta). \tag{A.25}$$

We can simplify the formula for $T(\gamma, \eta)$ as follows:

$$\begin{aligned} T(\gamma, \eta) &= \eta\left(-h\left(\frac{\gamma}{2\eta}\right) + \left[\log \eta - \frac{\gamma}{\eta}\log \gamma \right.\right. \\ &\quad \left.\left. - \frac{\eta-\gamma}{\eta}\log(\eta-\gamma)\right]\right) \\ &= \eta\left(\left[\left(\frac{\gamma}{\eta}\log\eta - \frac{\gamma}{\eta}\log\gamma\right)\right.\right. \\ &\quad \left.\left. + \left(\frac{\eta-\gamma}{\eta}\log\eta - \frac{\eta\gamma}{\eta}\log(\eta-\gamma)\right)\right] - h\left(\frac{\gamma}{2\eta}\right)\right] \\ &= \eta\left(h\left(\frac{\gamma}{\eta}\right) - h\left(\frac{\gamma}{2\eta}\right)\right). \end{aligned}$$

Since $0 \le \gamma \le \eta$ and the second factor depends only on $\rho \triangleq \gamma/\eta$,

it is clearly best to choose $\eta = 1$. Thus

$$C_q = \max_{0 \le \rho \le 1} \left(h(\rho) - h\left(\frac{\rho}{2}\right) \right). \quad \text{(A.26)}$$

Since $(d/d\rho)(h(\rho) - h(\rho/2)) = \log((1-\rho)/\rho) - (1/2)\log((2-\rho)/\rho) = 0$ implies $((1-\rho)/\rho)^2 = ((2-\rho)/\rho)$ or $\rho^2 - 2\rho + (1/2) = 0$, we get the maximum at $\rho = 1 - (1/2)\sqrt{2} \sim 0.293$. From a table for the binary entropy function $0.868 \le h(0.293) \le 0.875$, $0.597 \le h(0.1465) \le 0.603$, and

$$0.265 \le C_q \le 0.278, q = \left(\frac{1}{2}, \frac{1}{2}\right). \quad \text{(A.27)}$$

Additional calculations, which we omit, give also $C = C_q$ for $q = (1/2, 1/2)$.

Finally, we mention that it is tempting to conjecture that c) of Proposition 1 can be improved to c') $|\mathcal{U}| \le \min(|\mathcal{X}| + |\mathcal{S}|, |\mathcal{Y}|)$. If true, this would greatly simplify calculations of capacities in some cases.

C. Condition (T) does not imply condition (T_1)

Since Example 3 satisfies (T) it suffices to show that it does not satisfy (T_1). For this we restrict the minimization to distributions q, q' concentrated on single elements of $\mathcal{S}$ and show that for every $u, v : \mathcal{S} \to \mathcal{X}$

$$\min_{s, s' \in \mathcal{S}} \sum_y |w(y|u(s)|s) - w(y|v(s')|s')| = 0. \quad \text{(A.28)}$$

In fact, there are exactly three different vectors occurring as rows of the three matrices. On the other hand, for any $u : \mathcal{S} \to \mathcal{X}$ there are at least two different vectors among the vectors $w(\cdot|u(s)|s)$, $s \in \mathcal{S}$. Since $2 + 2 > 3$ for every $u, v : \mathcal{S} \to \mathcal{X}$ there exist $s, s' \in \mathcal{S}$ with $w(\cdot|u(s)|s) = w(\cdot|v(s')|s')$ and thus (28).

REFERENCES

[1] S. I. Gelfand and M. S. Pinsker, "Coding for channels with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[2] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie u. verw. Geb.*, vol. 44, pp. 159–175, 1978.

[3] ——, "Coloring hypergraphs: a new approach to multiuser source coding," Part I, *J. Combin. Inform. Syst. Sci.*, vol. 4, no. 1, pp. 76–115, 1979; Part II ib. vol. 5, no. 3, pp. 220–268, 1980.

[4] J. Kiefer and J. Wolfowitz, "Channels with arbitrarily varying channel probability functions," *Inform. Contr.*, vol. 5, pp. 44–54, 1962.

[5] A. V. Kusnetsov and B. S. Tsybakov, "Coding in a memory with defective cells," *Probl. Pered. Inform.*, vol. 10, no. 2, pp. 52–60, 1974.

[6] C. Heegard and 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.

[7] J. Wolfowitz, *Coding Theorems of Information Theory*, 3 ed. Berlin, Heidelberg, New York: Springer-Verlag, 1978.

[8] R. Ahlswede, "Geometric single-letterization techniques and capacities for AV-channels," in preparation.

[9] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Dev.*, vol. 2, pp. 289–293, 1958.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1982.