Correspondence

Zero-Error List Capacities of Discrete Memoryless Channels

İ. Emre Telatar, *Member, IEEE*

***Abstract*—We define zero-error list capacities for discrete memoryless channels. We find lower bounds to, and a characterization of these capacities. As is usual for such zero-error problems in information theory, the characterization is not generally a single-letter one. Nonetheless, we exhibit a class of channels for which a single letter characterization exists. We also show how the computational cutoff rate relates to the capacities we have defined.**

***Index Terms*—Acyclic channels, cutoff rate, list decoding, zero-error list capacity, zero undetected-error capacity.**

I. Introduction

It is sometimes desirable that the decoder of a communication system declare not just one, but several estimates of the transmitted data [1]. For example, the encoder and the decoder may be the inner code of a more complex transmission system, the structure of the outer code can then be used to choose among the estimates the inner code provides. Or, the data source that is driving the transmission system may have redundancy (which for some reason, e.g., delay considerations, has not been removed). This redundancy can be used at a later stage to pick one of the estimates. A decoder that may produce more than one estimate is called a *list decoder*. In this correspondence, we will investigate the performance of list decoders on discrete memoryless channels under a *zero-error* constraint.

Suppose we are given a discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $\{P(y|x),\ y \in \mathcal{Y},\ x \in \mathcal{X}\}$. The extension of the transition probability matrix to blocks of n inputs and outputs is denoted by P^n, and by the memoryless property for $x = (x_1, \cdots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \cdots, y_n) \in \mathcal{Y}^n$

$$P^n(y|x) = \prod_{i=1}^{n} P(y_i|x_i).$$

A block code of length n for a DMC with input alphabet $\mathcal{X}$ is a collection $\mathcal{C} \subset \mathcal{X}^n$ of sequences of input letters of length n. Elements of $\mathcal{C}$ are called codewords. A *zero-error list decoder* for a block code $\mathcal{C}$ is a decoder that assigns to every output $y \in \mathcal{Y}^n$ the set of codewords $\mathcal{L}(y, \mathcal{C}) \subset \mathcal{C}$ that could have produced that output with positive probability: $\mathcal{L}(y, \mathcal{C}) = \{c \in \mathcal{C}\colon P^n(y|c) > 0\}$. That is, the decoder decides on a list of codewords rather than a single codeword. It is clear that if a codeword c is transmitted and an output y is received, the transmitted codeword c always appears on the list (hence, the name "zero-error"), and that among the zero-error schemes this one produces the shortest list for any output y. Let $L(y, \mathcal{C}) = |\mathcal{L}(y, \mathcal{C})|$ be the size of the list. We assume that the codewords are equally probable, and define the ρth moment of the list size by

$$E[L^\rho] = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}^n} P^n(y|c) L(y, \mathcal{C})^\rho.$$

For $\epsilon > 0$ define $M(n, P, \rho, \epsilon)$ as the maximum size of codes of blocklength n such that the ρth moment of the list size is at most $1 + \epsilon$ when these codes are used over the DMC P. Now define the *zero-error* ρth-*moment list capacity* of a DMC P as

$$C_{0\ell}(\rho, P) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \ln M(n, P, \rho, \epsilon). \tag{1}$$

In the following, we will give lower bounds to $C_{0\ell}(\rho, P)$ and also a non-single-letter characterization of it. As is typical for such "singular" problems in information theory, no single-letter characterization of $C_{0\ell}(\rho, P)$ is known. Nonetheless, we will exhibit a nontrivial class of channels for which a single-letter characterization is possible. Furthermore, in the special case of $\rho \to \infty$, a single-letter characterization exists for all channels.

II. Characterization of $C_{0\ell}$

For $C_{0\ell}(\rho, P)$ to be positive, there must be an output which is not reachable from all inputs. Formally there must exist a triple $(x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$ such that

$$P(y|x_1) = 0 \quad \text{and} \quad P(y|x_2) > 0.$$

If there is no such triple, whatever the output word, all input words are possible and the decoder has to declare the entire codebook $\mathcal{C}$. Thus no rate larger than zero is possible. If, on the other hand, there is such a triple then $C_{0\ell} > 0$.

Theorem 1: For $\rho > 0$

$$C_{0\ell}(\rho, P) \geq \max_{Q} \min_{\substack{V, W\colon W \ll P \\ WQ = VQ}} I(Q, W) + \rho^{-1} D(V\|P|Q) \tag{2}$$

where Q ranges over the probability distributions on $\mathcal{X}$. Moreover, if we compute the lower bound for P^n, normalize, and pass to the limit

$$C_{0\ell}(\rho, P) = \lim_{n \to \infty} \frac{1}{n} \max_{Q} \min_{\substack{V, W\colon W \ll P^n \\ WQ = VQ}} I(Q, W) + \rho^{-1} D(V\|P^n|Q). \tag{3}$$

The notation $W \ll P$ means $W(y|x) = 0$ whenever $P(y|x) = 0$, (WQ) denotes the output distribution of the channel W when the input distribution is Q

$$D(V\|P|Q) = \sum_{x, y} Q(x) V(y|x) \ln \frac{V(y|x)}{P(y|x)}$$

and

$$I(Q, W) = \sum_{x, y} Q(x) W(y|x) \ln \frac{W(y|x)}{(WQ)(y)}$$

are the conditional divergence and the average mutual information, respectively.

We will give a proof of this theorem after we discuss some of its applications. Note that $C_{0\ell}(\rho, P)$ is nonincreasing in ρ. This is clear both from the definition and the formulation in Theorem 1.

Manuscript received January 10, 1996; revised March 25, 1997. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Whistler, BC, Canada, September 17–22, 1995.

The author is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA.

Publisher Item Identifier S 0018-9448(97)06706-0.

Remark: In the minimization over V and W in (2), we may add the constraint $V \ll P$ to the constraints $W \ll P$ and $WQ = VQ$. This is because if $V \not\ll P$ then there are only two possibilities:

1) for some x, y for which $Q(x) > 0$, $V(y|x) > P(y|x) = 0$, and thus $D(V\|P|Q) = \infty$;
2) for all x, y, $Q(x)P(y|x) = 0$ implies $Q(x)V(y|x) = 0$. In this case, we can replace V with V' defined as

$$V'(y|x) = \begin{cases} V(y|x), & \text{if } Q(x) > 0 \\ P(y|x), & \text{else.} \end{cases}$$

Then $V' \ll P$ and $Q \times V' = Q \times V$, thus the value of the objective function and the other constraints are not changed.

The case of $\rho = 1$ is of particular interest, the corresponding capacity $C_{0\ell}(1, P)$ is called the *zero-error average list size capacity*. Theorem 1 implies

$$C_{0\ell}(1, P) \geq \max_{Q} \min_{\substack{V, W: W \ll P \\ WQ = VQ}} I(Q, W) + D(V\|P|Q)$$

and

$$C_{0\ell}(1, P) = \lim_{n \to \infty} \frac{1}{n} \max_{Q} \min_{\substack{V, W: W \ll P^n \\ WQ = VQ}} I(Q, W) + D(V\|P^n|Q)$$

recovering the results of [2].

Another special case is obtained by letting ρ become vanishingly small. The constraint on the ρth moment of the list size is then equivalent to demanding that $\Pr[L > 1]$ gets arbitrarily small. As $\rho \to 0$, we see that in the minimization (2) V needs to be chosen so as to satisfy $D(V\|P|Q) = 0$, equivalently $V(y|x)Q(x) = P(y|x)Q(x)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and we get

$$C_{0\ell}(0^+, P) \geq \max_{Q} \min_{\substack{W: W \ll P \\ WQ = PQ}} I(Q, W)$$

recovering the previously known lower bound for *zero-undetected-error capacity* C_{0u} [2]–[5].

As a further special case, consider $\rho \to \infty$. Let us define

$$C_{0\ell}(\infty, P) \triangleq \lim_{\rho \to \infty} C_{0\ell}(\rho, P). \tag{4}$$

One might think that $C_{0\ell}(\infty, P)$ should equal the zero-error capacity C_0 as defined by Shannon [6] by arguing that demanding the ∞th power of $L(y, \mathcal{C})$ to be arbitrarily close to 1 is equivalent to demanding that $L(y, \mathcal{C})$ equal 1 with probability 1. This is not the case, because of the order we take limits: for any large but finite ρ, we can make the ρth moment of L decay to 1 without requiring that the probability of $L > 1$ equals zero. Surprisingly, one can give a single-letter expression for $C_{0\ell}(\infty, P)$.

Theorem 2:

$$C_{0\ell}(\infty, P) = \min_{W: W \ll P} C(W)$$

where $C(W) = \max_Q I(Q, W)$ denotes the ordinary capacity of a discrete memoryless channel W.

Proof: That

$$C_{0\ell}(\infty, P) \geq \min_{W: W \ll P} C(W)$$

follows from omitting the second term in (2) to obtain

$$C_{0\ell}(\infty, P) \geq \max_{Q} \min_{\substack{V, W: VQ = WQ \\ V \ll P, W \ll P}} I(Q, W)$$

observing that choosing $V = W$ enlarges the feasible set for W and thus

$$C_{0\ell}(\infty, P) \geq \max_{Q} \min_{W: W \ll P} I(Q, W)$$

and finally noting that I is concave in its first and convex in its second argument and that $\{W: W \ll P\}$ is a convex set thus concluding that the maximization and minimization can be interchanged to give

$$C_{0\ell}(\infty, P) \geq \min_{W: W \ll P} C(W).$$

We now need to show the converse inequality

$$C_{0\ell}(\infty, P) \leq \min_{W: W \ll P} C(W).$$

To do this, let $W^* \ll P$ be such that

$$C(W^*) = \min_{W: W \ll P} C(W).$$

Then, by choosing $V = W = W^{*n}$ in (3)

$$C_{0\ell}(\rho, P) \leq \lim_{n \to \infty} \max_{Q} n^{-1} I(Q, W^{*n}) + (n\rho)^{-1} D(W^{*n}\|P^n|Q)$$

Now observe that

$$n^{-1} D(W^{*n}\|P^n|Q) \leq \max_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} W^*(y|x) \ln (W^*(y|x)/P(y|x)) < \infty$$

and thus

$$\lim_{\rho \to \infty} C_{0\ell}(\rho, P) \leq \lim_{n \to \infty} \max_{Q} n^{-1} I(Q, W^{*n}) = C(W^*)$$

completing the proof. □

Remark: For a discrete memoryless channel P with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, for $y \in \mathcal{Y}$ define

$$S_y = \{x \in \mathcal{X}: P(y|x) > 0\}$$

and let

$$\pi_0 = \min_{Q} \max_{y \in \mathcal{Y}} \sum_{x \in S_y} Q(x).$$

It is known [6] that if the zero-error feedback capacity $C_{0f}(P)$ of the channel P is positive it equals $\ln(1/\pi_0)$. In [7], it is proved that $\ln(1/\pi_0) = \min_{W: W \ll P} C(W)$. Furthermore, in [8] it is proved that if $C_0(L, P)$ denotes the zero-error capacity of the channel P for a *fixed* list size L, then

$$\lim_{L \to \infty} C_0(L, P) = \ln(1/\pi_0)$$

(without the positivity condition). We thus see that

$$\lim_{\rho \to \infty} C_{0\ell}(\rho, P) = \lim_{L \to \infty} C_0(L, P) = C_{0f}(P)$$

where the second equality holds whenever $C_{0f}(P) > 0$.

We have thus seen that $C_{0\ell}(\infty, P)$ has a single-letter characterization. A more surprising result is that for a special class of DMC's one can obtain a single-letter expression for $C_{0\ell}(\rho, P)$ for any $\rho > 0$.

Theorem 3: Given a DMC P with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, construct the bipartite graph $G(P)$ with vertices $\mathcal{X} \cup \mathcal{Y}$ and edges

$$\{(x, y): x \in \mathcal{X}, y \in \mathcal{Y}, P(y|x) > 0\}.$$

If $G(P)$ is acyclic then

$$C_{0\ell}(\rho, P) = E_0(\rho, P)/\rho$$

where

$$E_0(\rho, P) = \max_{Q} -\ln \sum_{y} \left[\sum_{x} Q(x) P(y|x)^{1/(1+\rho)} \right]^{1+\rho}.$$

This result is similar to that of [9] where it is shown that for the same class of channels the zero-undetected-error capacity C_{0u} is equal to the ordinary capacity C.

Proof: We claim that for such channels

$$(V \ll P,\, W \ll P,\, VQ = WQ) \Longrightarrow Q \times V = Q \times W.$$

From this claim and the remark following Theorem 1 it follows that

$$\max_{Q} \min_{\substack{V,\, W:\, W \ll P \\ VQ = WQ}} I(Q,\, W) + \rho^{-1} D(V\|P|Q) = \max_{Q} \min_{W:\, W \ll P} I(Q,\, W) + \rho^{-1} D(W\|P|Q).$$

From [10, Prob. 23, pp. 192, 193] the expression on the right is equal to $E_0(\rho,\, P)/\rho$. Noting that $E_0(\rho,\, P^n) = nE_0(\rho,\, P)$ [11, Theorem 5], the proof follows.

It remains to prove the claim: Given $W \ll P$ and $V \ll P$ with $WQ = VQ$ note that

$$\sum_x Q(x)W(y|x) = \sum_x Q(x)V(y|x) \tag{5}$$

and

$$\sum_y Q(x)W(y|x) = \sum_y Q(x)V(y|x). \tag{6}$$

Suppose that there exist $x_0 \in \mathcal{X},\, y_0 \in \mathcal{Y}$, such that

$$Q_{x_0} W_{y_0|x_0} \neq Q_{x_0} V_{y_0|x_0}.$$

Then, to satisfy (5) there must exist $x_1 \neq x_0$ such that

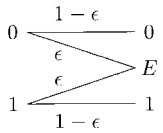
$$Q_{x_1} W_{y_0|x_1} \neq Q_{x_1} V_{y_0|x_1}.$$

To satisfy (6), there must exist $y_1 \neq y_0$ such that

$$Q_{x_1} W_{y_1|x_1} \neq Q_{x_1} V_{y_1|x_1}.$$

Continuing in this manner, we find a sequence $x_0,\, y_0,\, x_1,\, y_1,\, \cdots$ such that $x_n \neq x_{n+1},\, y_n \neq y_{n+1},\, Q_{x_n} W_{y_n|x_n} \neq Q_{x_n} V_{y_n|x_n}$, and $Q_{x_{n+1}} W_{y_n|x_{n+1}} \neq Q_{x_{n+1}} V_{y_n|x_{n+1}}$. The inequalities imply that at least one of $W_{y_n|x_n}$ and $V_{y_n|x_n}$ and at least one of $W_{y_n|x_{n+1}}$ and $V_{y_n|x_{n+1}}$ must be positive. Since $V \ll P$ and $W \ll P$, we conclude that $P(y_n|x_n) > 0$ and $P(y_n|x_{n+1}) > 0$. Furthermore, $x_0,\, x_1,\, \cdots$ must be all distinct, otherwise, if say $x_n = x_{n+m}$ then the sequence of nodes $x_n,\, y_n,\, \cdots,\, x_{n+m}$ would form a cycle in $G(P)$. Since $|\mathcal{X}|$ is finite, this is a contradiction. □

Example (Binary Erasure Channel): Consider a channel with a binary input $\mathcal{X} = \{0,\, 1\}$ and ternary output $\mathcal{Y} = \{0,\, 1,\, E\}$ with transition probabilities as below:



The channel is clearly acyclic, and thus

$$C_{0\ell}(\rho,\ \text{BEC}) = E_0(\rho,\ \text{BEC})/\rho = -\rho^{-1} \ln\left(\epsilon + (1-\epsilon)/2^{\rho}\right).$$

It is instructive to compare the zero-error list capacity $C_{0\ell}$ to its nonzero-error counterpart. To that end, for $c \in \mathcal{C}$, let $\mathcal{L}_c(y,\, c,\, \mathcal{C})$ be the set of codewords in $\mathcal{C}$ whose likelihood is at least as great as that of c, when y is received, i.e.,

$$\mathcal{L}_c(y,\, c,\, \mathcal{C}) = \{k \in \mathcal{C}\colon P^n(y|k) \geq P^n(y|c)\}.$$

Let $L_c(y,\, c,\, \mathcal{C}) = |\mathcal{L}_c(y,\, c,\, \mathcal{C})|$ be the number of codewords which are as likely as c when y is received. We will assume that the codewords are equally probable and we will define the ρth moment of the number of codewords as likely as the transmitted codeword by

$$E[L_c^{\rho}] = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}^n} P^n(y|c) L_c(y,\, c,\, \mathcal{C})^{\rho}.$$

For $\epsilon > 0$, define $M_c(n,\, P,\, \rho,\, \epsilon)$ as the maximum size of codes with blocklength n such that $E[L_c^{\rho}] < 1 + \epsilon$. Now let

$$C_{\ell}(\rho,\, P) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \ln M_c(n,\, P,\, \rho,\, \epsilon).$$

An equivalent way of thinking about C_{ℓ} is as follows. Suppose we have a decoder aided by a genie that answers the questions of the form "is c the correct codeword?" Let $G(c)$ be the random variable whose value is the number of questions the decoder needs to ask the genie until it is answered in the affirmative. $C_{\ell}(\rho,\, P)$ is the highest rate for which the ρth moment of $G(c)$ can be made arbitrarily close to 1 for all codewords c. $C_{\ell}(1,\, P)$ is known as the *cutoff rate* of the channel P. It is clear that $C_{\ell}(\rho,\, P) \geq C_{0\ell}(\rho,\, P)$.

Theorem 4. ([12]–[17]): For $\rho > 0$

$$C_{\ell}(\rho,\, P) = E_0(\rho)/\rho = \max_{Q} \min_{W} I(Q,\, W) + \rho^{-1} D(W\|P|Q). \tag{7}$$

The formal similarity of (2) and (7) is remarkable.

Corollary 1: For the channels described in Theorem 3

$$C_{0\ell}(\rho,\, P) = C_{\ell}(\rho,\, P).$$

We will now prove Theorem 1.

Converse Part of Theorem 1: Suppose we are given a code $\mathcal{C} \subset \mathcal{X}^n$ of rate R and that the ρth moment of $L(y,\, \mathcal{C})$ is less than $1 + \epsilon$

$$|\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} \sum_{y \in \mathcal{Y}^n} P^n(y|c) L(y,\, \mathcal{C})^{\rho} < 1 + \epsilon.$$

Now let $\mathcal{D} = \{y \in \mathcal{Y}^n\colon L(y,\, \mathcal{C}) > 0\}$ and choose a distribution Q on $\mathcal{X}^n$ with

$$Q(x) = \begin{cases} 1/|\mathcal{C}|, & \text{if } x \in \mathcal{C} \\ 0, & \text{else.} \end{cases}$$

Fix any two auxiliary channels $V \ll P^n$ and $W \ll P^n$ with $WQ = VQ$. Let $\hat{W}$ be the reverse channel

$$\hat{W}(x|y) = Q(x)W(y|x)/(WQ)(y).$$

Then,

$$\begin{aligned}
&\ln(1+\epsilon) \\
&\quad > \ln\left(\sum_{x \in \mathcal{X}^n} \sum_{y \in \mathcal{D}} Q(x) P^n(y|x) L(y,\, \mathcal{C})^{\rho}\right) \\
&\quad = \ln\left(\sum_{x \in \mathcal{X}^n} \sum_{y \in \mathcal{D}} Q(x) V(y|x) \frac{P^n(y|x) L(y,\, \mathcal{C})^{\rho}}{V(y|x)}\right) \\
&\quad \geq -\sum_{x \in \mathcal{X}^n} \sum_{y \in \mathcal{D}} Q(x) V(y|x) \ln \frac{V(y|x)}{P^n(y|x) L(y,\, \mathcal{C})^{\rho}} \\
&\quad = -D(V\|P^n|Q) + \rho \sum_{x \in \mathcal{X}^n} \sum_{y \in \mathcal{D}} Q(x) V(y|x) \ln L(y,\, \mathcal{C}) \\
&\quad = -D(V\|P^n|Q) + \rho \sum_{x \in \mathcal{X}^n} \sum_{y \in \mathcal{D}} Q(x) W(y|x) \ln L(y,\, \mathcal{C}) \\
&\quad \geq -D(V\|P^n|Q) + \rho(H(Q) - I(Q,\, W)) \\
&\quad = -D(V\|P^n|Q) - \rho I(Q,\, W) + \rho n R
\end{aligned}$$

and thus we obtain

$$\frac{1}{n}\left(\rho^{-1} D(V\|P^n|Q) + I(Q,\, W)\right) \geq R - \frac{1}{n\rho} \ln(1+\epsilon)$$

proving the converse. Note that we have proved more than we claimed. Namely, for any positive sequence $\{\epsilon_n\}_{n \geq 1}$ with

$$\limsup_{n \to \infty} (1/n) \ln(1 + \epsilon_n) = 0$$

and for any $\rho > 0$

$$\limsup_{n\to\infty} (1/n) \ln M(n, P, \rho, \epsilon_n) \le C_{0\ell}(\rho, P).$$

In particular, for rates above $C_{0\ell}(\rho, P)$ the ρth moment of the list size grows to infinity exponentially in the blocklength n. □

To prove the direct part of Theorem 1, we will need some preliminaries. For $x = (x_1, \cdots, x_n) \in \mathcal{X}^n$ let $T[x]$ denote its type; $T[x]$ is a probability distribution on $\mathcal{X}$ with

$$T[x](u) = \frac{1}{n} |\{k: x_k = u\}|.$$

For an input distribution Q, let $\mathcal{T}_Q^n \subset \mathcal{X}^n$ denote the set of all x with $T[x] = Q$. For $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$ let $T[x, y]$ denote the joint type of x and y. $T[x, y]$ is a probability distribution on $\mathcal{X} \times \mathcal{Y}$ with

$$T[x, y](u, v) = \frac{1}{n} |\{k: x_k = u, y_k = v\}|.$$

Note that if $T[x] = Q$ then $T[x, y]$ is necessarily of the form

$$T[x, y](u, v) = Q(u)V(v|u)$$

for some conditional distribution V. If $T[x] = Q$ we let

$$T[y|x] = V$$

to mean $T[x, y] = Q \times V$, and say that y has conditional type V with respect to x.

Lemma 1: Let V and W be conditional types. Let (C_1, C_2, Y) be a random variable on $\mathcal{X}^n \times \mathcal{X}^n \times \mathcal{Y}^n$ with probability distribution

$$\Pr(C_1 = c_1, C_2 = c_2, Y = y) = \begin{cases} |\mathcal{T}_Q^n|^{-2} P^n(y|c_1), & \text{if } c_1 \in \mathcal{T}_Q^n \text{ and } c_2 \in \mathcal{T}_Q^n \\ 0, & \text{else.} \end{cases}$$

Then

$$\Pr\{T[Y|C_1] = V\} \le \exp[-nD(V\|P|Q)] \tag{8}$$

and

$$\Pr\{T[Y|C_2] = W \,|\, T[Y|C_1] = V\} \simeq \begin{cases} \exp[-nI(Q, W)], & \text{if } WQ = VQ \\ 0, & \text{else} \end{cases} \tag{9}$$

where we use the notation $a(n) \simeq b(n)$ to mean

$$\lim_{n\to\infty} (1/n) \ln a(n) = \lim_{n\to\infty} (1/n) \ln b(n)$$

that is, if two codewords are chosen independently and uniformly from $\mathcal{T}_Q^n$ and the first is transmitted, then the conditional type of the received sequence with respect to the transmitted codeword equals V with probability $\exp[-nD(V\|P|Q)]$ and the conditional type of this received sequence with respect to the other codeword equals W with conditional probability $\exp[-nI(Q, W)]$.

Proof: Let $a \in \mathcal{X}^n$ and $b \in \mathcal{Y}^n$ be such that $T[x, y] = Q \times V$. Then

$$P^n(b|a) = \prod_{k=1}^n P(b_k|a_k) = \prod_{x,y} P(y|x)^{nQ(x)V(y|x)} = V^n(b|a) \exp[-nD(V\|P|Q)].$$

Thus

$$\begin{aligned} \Pr\{T[Y|C_1] = V\} &= \sum_{a,b:\, T[b|a]=V} \Pr\{(C_1, Y) = (a, b)\} \\ &= \sum_{a,b:\, T[b|a]=V} \Pr\{C_1 = a\} P^n(b|a) \\ &\le \sum_{a,b} \Pr\{C_1 = a\} V^n(b|a) \exp[-nD(V\|P|Q)] \\ &= \exp[-nD(V\|P|Q)] \end{aligned}$$

proving the first part of the lemma. To prove the second part, first observe that since $T[C_1] = T[C_2] = Q$, $T[y] = WQ$, and $T[y] = VQ$. Thus if $VQ \neq WQ$, then no triple (c_1, c_2, y) satisfies $T[y|c_1] = V$ and $T[y|c_2] = W$ and hence

$$\Pr\{T[Y|C_2] = W \,|\, T[Y|C_1] = V\} = 0.$$

If, on the other hand, $VQ = WQ$, let for $y \in \mathcal{T}_{VQ}^n$

$$A(y) = \{x \in \mathcal{T}_Q^n : T[y|x] = W\}.$$

Then

$$\begin{aligned} \Pr\{T[Y|C_2] = W \,|\, T[Y|C_1] = V\} &= \Pr\{C_2 \in A(y)\} \\ &= |A(y)|/|\mathcal{T}_Q^n|. \end{aligned}$$

The size of $A(y)$ given by

$$|A(y)| = \prod_v \frac{(nT[y](v))!}{\prod_u (nQ(u)W(v|u))!}$$

is independent of y and

$$|A(y)| \simeq |\mathcal{T}_Q^n| \exp[-nI(Q, W)]$$

proving the second part of the lemma. □

Another result we will need is about the sums of independent 0–1 random variables:

Lemma 2: Given $\rho > 0$, $r > 0$, $\alpha \ge 0$. For $n = 1, 2, \cdots$, let

$$S_n = 1 + B(m_n, p_n)$$

where $B(m, p)$ is a binomial random variable with parameters m and p, $m_n = \lceil \exp(nr) \rceil$, and $p_n = \exp(-n(\alpha - o(n)))$ with $\lim_{n\to\infty} o(n) = 0$. Then

$$E[S_n^\rho] \le \begin{cases} 1 + o'(n), & \text{if } r < \alpha \\ \exp(n\rho(r - \alpha + o'(n))), & \text{if } r \ge \alpha \end{cases}$$

where $o'(n)$ satisfies $\lim_{n\to\infty} o'(n) = 0$. These two inequalities can be summarized as

$$E[S_n^\rho] \le 1 + o'(n) + \exp(n\rho(r - \alpha + o'(n))).$$

Proof: We will consider the two cases indicated in the lemma:

1) $r < \alpha$: Consider the moment generating function of S_n

$$\phi_{S_n}(\lambda) = E[\exp(\lambda S_n)] = e^\lambda (1 + p_n(e^\lambda - 1))^{m_n}.$$

Since $\lim m_n = \infty$ and $\lim m_n p_n = 0$, the moment generating function of S_n tends to that of a random variable that takes the value 1 with probability 1. Thus we conclude that [18, p. 408]

$$E[S_n^\rho] = 1 + o_1(n).$$

with

$$\lim_{n\to\infty} o_1(n) = 0.$$

2) $r \ge \alpha$: Let $\tilde{p}_n = p_n e^{n\delta_n}$ where δ_n is chosen such that

$$\lim \delta_n = 0 \qquad p_n \le \tilde{p}_n \le 1 \quad \text{and} \qquad \lim m_n \tilde{p}_n = \infty.$$

Such a choice always exists; for example, one can take

$$\delta(n) = \begin{cases} 0, & \text{if } r > \alpha \\ \min\{|o(n)| + 1/\sqrt{n}, \alpha - o(n)\}, & \text{if } r = \alpha. \end{cases}$$

Let $\tilde{S}_n = 1 + B(m_n, \tilde{p}_n)$. Clearly,

$$E[S_n^\rho] \le E[\tilde{S}_n^\rho] = (m_n \tilde{p}_n)^\rho E\left[\left(\frac{\tilde{S}_n}{m_n \tilde{p}_n}\right)^\rho\right].$$

As n gets large, $m_n\tilde{p}_n$ tends to ∞ and the moment generating function of $\tilde{S}_n/(m_n\tilde{p}_n)$

$$\phi(\lambda) = e^{\lambda/(m_n\tilde{p}_n)}\left[1 + p_n\left(e^{\lambda/(m_n\tilde{p}_n)} - 1\right)\right]^{m_n}$$

tends to e^{λ}, that of random variable that takes the value 1 with probability 1. We conclude that

$$\begin{aligned} E[S_n^{\rho}] &\le E[\tilde{S}_n^{\rho}] \\ &= (m_n\tilde{p}_n)^{\rho}(1 + o_2(n)) \\ &\le \exp(n\rho(r - \alpha + o(n) + \delta(n)))\exp(o_2(n)) \\ &= \exp(n\rho(r - \alpha + o_3(n))), \end{aligned}$$

where $\lim o_2(n) = 0$ and $o_3(n) = o_2(n)/(n\rho) + o(n) + \delta(n)$ also satisfies $\lim o_3(n) = 0$.

Letting $o'(n) = \max\{o_1(n), o_3(n)\}$ completes the proof. □

Direct Part of Theorem 1: Consider an ensemble of codes of blocklength n with $M = \lceil \exp(nR) \rceil$ codewords where the codewords C_k are chosen independently and according to a uniform distribution over $\mathcal{T}_Q^n$, the set of Q-typical sequences of length n. We will upper-bound $E[L(y, \mathcal{C})^{\rho}]$.

Without loss of generality, suppose the first codeword is transmitted. The probability space we have is then

$$\mathcal{X}^n \times \cdots \times \mathcal{X}^n \times \mathcal{Y}^n$$

with the probability measure

$$\Pr\{(C_1, \cdots, C_M, Y) = (c_1, \cdots, c_M, y)\} = \begin{cases} |\mathcal{T}_Q^n|^{-M} P^n(y|c_1), & \text{if } \forall k\ c_k \in \mathcal{T}_Q^n \\ 0, & \text{else.} \end{cases}$$

From Lemma 1, we know that

$$\Pr[T[Y|C_1] = V] \le \exp[-nD(V\|P|Q)].$$

Conditional on $T[Y|C_1] = V$, the probability $p(V)$ that $P^n(Y|C_2) > 0$ is given by (again by Lemma 1, by summing over W and noting that there are polynomially many distinct conditional types)

$$p(V) = \exp\left\{-n[\min_{\substack{W:\ W \ll P \\ VQ = WQ}} I(Q, W) - o(n)]\right\}$$

where $\lim_{n\to\infty} o(n) = 0$. For $i \ge 2$, let X_i be the indicator random variable of the event $\{P^n(Y|C_i) > 0\}$. Conditional on $T[Y|C_1] = V$, $X_2, \cdots, X_M$ are independent, identically distributed 0–1 random variables with mean $p(V)$. Furthermore, the list size L is given by

$$L = 1 + \sum_{i=2}^{M} X_i.$$

Using Lemma 2 with

$$\alpha = \min_{\substack{W:\ W \ll P \\ VQ = WQ}} I(Q, W)$$

and $r = R$, we conclude that

$$E[L^{\rho} \mid T[Y|C_1] = V] \le 1 + o'(n) + \exp\{-n\rho(\min_{\substack{W:\ W \ll P \\ VQ = WQ}} I(Q, W) - R - o'(n))\}$$

where $\lim_{n\to\infty} o'(n) = 0$. Removing the conditioning by multiplying by the probability $\Pr\{T[Y|C_1] = V\}$ and summing over V (and noting again that there are only polynomially many distinct conditional types) we see that

$$\begin{aligned} E[L^{\rho}] &\le 1 + o'(n) + \exp\{-n \min_{V}[D(V\|P|Q) \\ &\qquad + \rho(\min_{\substack{W:\ W \ll P \\ VQ = WQ}} I(Q, W) - R) - o''(n)]\} \\ &= 1 + o'(n) + \exp\{-n[\min_{\substack{V, W:\ W \ll P \\ WQ = VQ}} [D(V\|P|Q) \\ &\qquad + \rho I(Q, W)] - \rho R - o''(n)]\}. \end{aligned}$$

Now observe that for all R less than the right-hand side of (2) the exponential term decays to zero with increasing n, proving (2). Applying (2) to P^n instead of P we complete the proof of the theorem. □

III. Conclusion

For zero-error list decoding we find achievable rates for which the ρth moment of the list size remains bounded. We give a single-letter lower bound for the capacity and also a non-single-letter characterization of it. We show that in the limit as ρ tends to infinity, the capacity can be found by a single-letter expression. We demonstrate that for acyclic channels the capacity has a single-letter characterization. We also show how the computation cutoff rate is related to the quantities investigated in this correspondence.

References

[1] G. D. Forney, "Exponential error bounds for erasure, list and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.

[2] R. Ahlswede, N. Cai, and Z. Zhang, "Erasure, list and detection zero-error capacities for low noise and a relation to identification," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 55–62, Jan. 1996.

[3] İ. E. Telatar and R. G. Gallager, "Zero error decision feedback capacity of discrete memoryless channels," in *Proc. 1990 Bilkent Int. Conf. on New Trends in Communication, Control, and Signal Processing* (Bilkent University, Ankara, Turkey, July 1990).

[4] İ. E. Telatar, "Multi-access communications with decision feedback decoding," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, May 1992.

[5] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," in *Proc. 1994 IEEE Int. Symp. on Information Theory* (Norwegian Institute of Technology, Trondheim, Norway, June 27–July 1 1994).

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

[7] R. Ahlswede, "Channels with arbitrarily varying channel probability functions in the presence of noiseless feedback," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 25, pp. 239–252, 1973.

[8] P. Elias, "Zero error capacity under list decoding," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1070–1074, Sept. 1988.

[9] M. S. Pinsker and A. Y. Sheverdyaev, "Transmission capacity with zero error and erasure," *Probl. Inform. Transm. (Probl. Pered. Inform.)*, vol. 6, no. 1, pp. 20–24, 1970.

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

[11] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.

[12] J. E. Savage, "Sequential decoding: The computation problem," *Bell Syst. Tech. J.*, vol. 45, pp. 149–175, 1966.

[13] I. M. Jacobs and E. R. Berlekamp, "A lowerbound to the distribution of computation for sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 167–174, Apr. 1967.

[14] F. Jelinek, "An upper bound on moments of sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 140–149, Jan. 1969.

[15] D. D. Falconer, "A hybrid coding scheme for discrete memoryless channels," *Bell Syst. Tech. J.*, vol. 48, pp. 691–728, 1969.

[16] E. Arıkan, "An upper bound on the cutoff rate of sequential decoding," *IEEE Trans. Inform. Theory*, vol. 34, pp. 55–63, Jan. 1988.

[17] E. Arıkan, "Lower bounds to moments of list size," in *IEEE Int. Symp. on Information Theory* (Abstracts of Papers), (San Diego, CA, Jan. 1990), pp. 145–146.
[18] P. Billingsley, *Probability and Measure*, 2nd ed. New York: Wiley, 1986.

A Non-Shannon-Type Conditional Inequality of Information Quantities

Zhen Zhang, *Senior Member, IEEE*, and
Raymond W. Yeung, *Senior Member, IEEE*

***Abstract*— Given n discrete random variables $\Omega = \{X_1, \cdots, X_n\}$, associated with any subset α of $\{1, 2, \cdots, n\}$, there is a joint entropy $H(X_\alpha)$ where $X_\alpha = \{X_i: i \in \alpha\}$. This can be viewed as a function defined on $2^{\{1, 2, \cdots, n\}}$ taking values in $[0, +\infty)$. We call this function the entropy function of Ω. The nonnegativity of the joint entropies implies that this function is nonnegative; the nonnegativity of the conditional joint entropies implies that this function is nondecreasing; and the nonnegativity of the conditional mutual informations implies that this function is two-alternative. These properties are the so-called basic information inequalities of Shannon's information measures. An entropy function can be viewed as a $2^n - 1$-dimensional vector where the coordinates are indexed by the subsets of the ground set $\{1, 2, \cdots, n\}$. As introduced in [4], Γ_n stands for the cone in $\mathbb{R}^{2^n-1}$ consisting of all vectors which have all these properties. Let Γ_n^* be the set of all $2^n - 1$-dimensional vectors which correspond to the entropy functions of some sets of n discrete random variables. A fundamental information-theoretic problem is whether or not $\overline{\Gamma}_n^* = \Gamma_n$. Here $\overline{\Gamma}_n^*$ stands for the closure of the set Γ_n^*. In this correspondence, we show that $\overline{\Gamma}_n^*$ is a convex cone, $\Gamma_2^* = \Gamma_2$, $\Gamma_3^* \neq \Gamma_3$, but $\overline{\Gamma}_3^* = \Gamma_3$. For four random variables, we have discovered a conditional inequality which is not implied by the basic information inequalities of the same set of random variables. This lends an evidence to the plausible conjecture that $\overline{\Gamma}_n^* \neq \Gamma_n$ for $n > 3$.**

***Index Terms*— Entropy, I-Measure, information inequalities, mutual information.**

I. Introduction and Summary

Let $\Omega_n = \{X_i: i = 1, \cdots, n\}$ be n jointly distributed discrete random variables with finite entropies. The basic Shannon's information measures associated with these random variables include all joint entropies, all conditional entropies, all mutual informations, and all conditional mutual informations involving some of these random variables. For any subset α of $\mathcal{N}_n = \{1, \cdots, n\}$ let

$$X_\alpha = \{X_i: i \in \alpha\}. \tag{1}$$

Let X_ϕ, where ϕ is the empty set, be a random variable taking a fixed value with probability 1. Define

$$I(\alpha, \beta|\gamma) = I(X_\alpha; X_\beta|X_\gamma). \tag{2}$$

Manuscript received October 30, 1995; revised February 15, 1997. This work was supported in part by the National Science Foundation under Grant NCR-9508282.

Z. Zhang is with the Communication Sciences Institute, Department of Electrical Engineering Systems, University of Southern California, Los Angeles, CA 90089-2565.

R. W. Yeung is with the Department of Information Engineering, The Chinese University of Hong Kong, NT, Hong Kong.

Publisher Item Identifier S 0018-9448(97)07295-7.

We see that when $\alpha = \beta$

$$I(\alpha, \alpha|\gamma) = H(X_\alpha|X_\gamma) \tag{3}$$

which is the conditional entropy; when $\gamma = \phi$

$$I(\alpha, \beta|\phi) = I(X_\alpha; X_\beta) \tag{4}$$

which is the unconditional mutual information, and when $\alpha = \beta$ and $\gamma = \phi$

$$I(\alpha, \alpha|\phi) = H(X_\alpha) \tag{5}$$

which is the joint entropy. This means that the function $I(\alpha, \beta|\gamma)$ covers all the basic Shannon's information measures. In this correspondence, all logarithms are in base 2.

It is well known that Shannon's information measures satisfy the following inequalities.

Proposition 1: For any three subsets α, β, and γ of $\mathcal{N}_n$, any set of n jointly distributed random variables X_i, $i = 1, \cdots, n$, with finite entropies

$$I(\alpha, \beta|\gamma) \geq 0. \tag{6}$$

These inequalities are called the *basic inequalities* of Shannon's information measures [4].

Let $H(\alpha) = I(\alpha, \alpha|\phi)$ be the joint entropy function. For any set of n jointly distributed random variables X_i, $i = 1, \cdots, n$, the associated entropies $H(\alpha)$ can be viewed as a function defined on $2^{\mathcal{N}_n}$

$$H: 2^{\mathcal{N}_n} \to [0, \infty). \tag{7}$$

The goal of this correspondence is to study this function for all possible sets of n random variables with finite entropies.

All basic Shannon's information measures can be expressed as linear functions of the joint entropies. Actually, we have

$$I(\alpha, \beta|\gamma) = H(\alpha \cup \gamma) + H(\beta \cup \gamma) - H(\alpha \cup \beta \cup \gamma) - H(\gamma). \tag{8}$$

The basic inequalities can be interpreted as a set of inequalities for the entropy function as follows.

Proposition 2: For any set of n jointly distributed random variables X_i, $i = 1, \cdots, n$, with finite entropies, the entropy function H associated with these random variables has the following properties.

1) For any two subsets α and β of $\mathcal{N}_n$

$$H(\alpha \cup \beta) + H(\alpha \cap \beta) \leq H(\alpha) + H(\beta). \tag{9}$$

Functions having this property are called two-alternative functions.

2) $\alpha \subset \beta$ implies

$$H(\alpha) \leq H(\beta). \tag{10}$$

Functions satisfying this property are called monotone nondecreasing, and

$$H(\phi) = 0. \tag{11}$$

It is easily seen from (8) that the first property corresponds to the nonnegativity of all mutual informations and condition mutual informations, and the second and third property correspond to the nonnegativity of all entropies and conditional entropies.