

# On Channel Capacity per Unit Cost

SERGIO VERDÚ

**Abstract**—Memoryless communication channels with arbitrary alphabets where each input symbol is assigned a cost are considered. The maximum number of bits that can be transmitted reliably through the channel per unit cost is studied. It is shown that if the input alphabet contains a zero-cost symbol, then the capacity per unit cost admits a simple expression as the maximum normalized divergence between two conditional output distributions. The direct part of this coding theorem admits a constructive proof via Stein's lemma on the asymptotic error probability of binary hypothesis tests. Single-user, multiple-access and interference channels are studied.

## I. INTRODUCTION

IN THE SHANNON THEORY, channel input constraints are usually modeled by a cost function that associates a nonnegative number  $b[x]$  to each element  $x$  of the input alphabet. The maximum number of bits per symbol that can be transmitted over the channel with average cost per symbol not exceeding  $\beta$  is the capacity-cost function,  $C(\beta)$ . Thus, the transmission of one bit of information requires  $1/C(\beta)$  symbols at cost equal to  $\beta/C(\beta)$ . Since the capacity-cost function is concave and nondecreasing, by varying  $\beta$  we can trade off the number of symbols and the cost it takes to send every unit of information through the channel. For example, Fig. 1 depicts the cost-per-bit as a function of the number of symbols-per-bit for a discrete-time additive white Gaussian channel with noise variance equal to  $\sigma^2$  and  $b[x] = x^2$ , i.e.,  $C(\beta) = \frac{1}{2} \log(1 + \beta/\sigma^2)$ . In this figure we see that the cost per bit escalates rapidly with the transmission rate if more than, say, one bit is to be transmitted per symbol, whereas the cost per bit is close to its infimum, viz.,  $\sigma^2 \ln 4$ , for, say, two or more symbols-per-bit. It may appear at first glance that the penalty for transmitting at a cost-per-bit close to its minimum value is, invariably, slow communication. However, exactly the opposite may be true. For example, consider the case when the discrete-time Gaussian channel arises in the analysis of the infinite-bandwidth continuous-time white Gaussian channel with noise power spectral density equal to  $\sigma^2 = N_0/2$  and input power limited to  $P$  energy units per second. In this context a symbol or channel use can be viewed as a degree of freedom or brief use of a narrow frequency

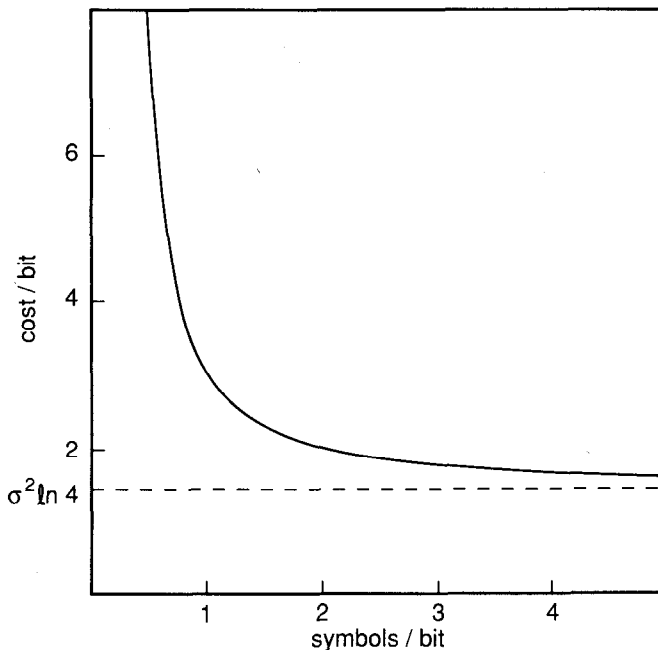


Fig. 1. Cost vs. number of symbols required to transmit one bit of information through AWGN channel.

band. Using a large number of degrees of freedom per transmitted bit allows us to spend as little energy as possible per bit, and thus, as little time as possible because any waveform whose energy is equal to  $E$  can be transmitted in  $E/P$  seconds. Thus, the capacity in bits per second of the continuous-time channels is simply  $P$  divided by the minimum energy-per-bit  $\sigma^2 \ln 4 = N_0 \ln 2$ , i.e., the well-known  $(P \log_2 e)/N_0$  bits/s.

The minimum cost incurred by the transmission of one bit of information through the channel is a fundamental figure of merit characterizing the most economical way to communicate reliably. Its reciprocal, the capacity per unit cost, is defined similarly to the conventional capacity, except that the ratio of the logarithm of the number of codewords to their blocklength (rate) is replaced by the ratio of the logarithm of the number of codewords to their cost (rate-per-unit cost).

Information-cost ratios have been studied, in various guises, in the past. Reza [19] considered the case in which the letters of the input alphabet need not have the same duration (studied by Shannon [21] in the context of noiseless source coding). Then the cost function  $b[x]$  becomes the duration of the symbol  $x$  in seconds and the rate of transmission in units per second is just the rate per unit

Manuscript received September 18, 1989. This work was supported in part by the Office of Naval Research under Contract N00014-87-K-0054 and Grant N00014-90-J-1734; and in part by the National Science Foundation under PYI Grant ECSE-8857689.

The author is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

IEEE Log Number 9036192.

cost. Motivated by this generalization, [12] and [16] gave algorithms for the computation of the capacity per unit cost of memoryless channels with finite input and output alphabets, under the restriction that  $\min_x b[x] > 0$ . Pierce [17] and subsequent works have found the capacity per unit cost in bits per joule (or in bits per photon) of various versions of the photon counting channel. The problem of finding the capacity (in units per second) of continuous-time channels with unlimited degrees of freedom (such as infinite-bandwidth Gaussian channels [1] and fading dispersive channels [9]) is equivalent, as we have illustrated, to finding the capacity per unit cost of an equivalent discrete-time channel. The first observation that in the Gaussian channel the minimum energy per bit is equal to  $N_0 \ln 2$  is apparently due to Golay [11]. In his seminal contribution [10], Gallager has abstracted the key features of channels where the main limitation is on the input cost rather than on the number of degrees of freedom and has investigated the reliability function of the rate per unit cost for discrete memoryless channels.

As we show in this paper, the capacity per unit cost can be computed from the capacity-cost function by finding  $\sup_{\beta > 0} C(\beta)/\beta$ , or, alternatively, as

$$C = \sup_x \frac{I(X;Y)}{E[b[X]]}. \quad (1)$$

However, the main contribution of this paper is to show that in the important case where the input alphabet contains a zero-cost symbol (which we label as "0") the capacity per unit cost is given by

$$C = \sup_x \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b[x]} \quad (2)$$

where the supremum is over the input alphabet,  $P_{Y|X=x}$  denotes the conditional output distribution given that the input is  $x$ , and  $D(P\|Q)$  is the (Kullback-Leibler) divergence between probability distributions  $P$  and  $Q$  (e.g., [2], [6]; also known as discrimination, relative entropy, information number, etc.). It turns out that (2) is much simpler to compute not only than (1) but also than the conventional capacity; the reason being twofold: the divergence between two conditional output distributions is usually easier to compute than the mutual information, and the required maximization is over the input alphabet itself rather than over the set of probability distributions defined on it. We illustrate this fact by several examples where even though the capacity is unknown, a simple evaluation of (2) results in a closed-form expression for the capacity per unit cost.

If the input alphabet is the real line and the cost function is  $b[x] = x^2$ , then the capacity per unit cost is lower bounded by one half of Fisher's information of the family of conditional output distributions. This bound is attained by additive non-Gaussian noise channels where the noise density function has heavier tails than the Gaussian. Moreover, using the Cramer-Rao bound we get an upper bound on the minimum energy required to

send one unit of information through the channel in terms of the minimum variance of an estimate of the input given the output.

We give a constructive proof of the achievability of (2) by pulse position modulation codes, via a simple application of Stein's lemma [4], a large-deviations result on the asymptotic probability of error of binary hypothesis tests, which is closely related to Shannon's source coding theorem. Although the definition of achievable rate per unit cost places no restrictions on the available blocklength, it turns out that codes whose blocklengths grow logarithmically with the number of codewords are sufficient to achieve capacity per unit cost.

We also study the capacity region per unit cost of memoryless multiple-access and interference channels. If each input alphabet contains a zero-cost symbol 0 then each user can transmit simultaneously at the capacity per unit cost of the single-user channel that results by having all other users transmit 0. This results in an easy-to-compute inner bound to the capacity region per unit cost. A sufficient condition (essentially, that the zero-cost symbol produces the most benign interference to other users) guarantees the optimality of that inner bound.

## II. CAPACITY PER UNIT COST OF SINGLE-USER CHANNELS

In this section we deal with single-user discrete-time channels without feedback and with arbitrary input and output alphabets denoted by  $A$  and  $B$ , respectively. An  $(n, M, \nu, \epsilon)$  code is one in which the blocklength is equal to  $n$ ; the number of codewords is equal to  $M$ ; each codeword  $(x_{m1}, \dots, x_{mn})$ ,  $m = 1, \dots, M$ , satisfies the constraint

$$\sum_{i=1}^n b[x_{mi}] \leq \nu$$

where  $b: A \rightarrow \mathbf{R}^+ = [0, +\infty)$  is the function that assigns a cost to each input symbol; and the average (over the ensemble of equiprobable messages) probability of decoding the correct message is better than  $1 - \epsilon$ .

We will adopt the following standard definition [6].

*Definition 1:* Given  $0 \leq \epsilon < 1$  and  $\beta > 0$ , a nonnegative number  $R$  is an  $\epsilon$ -achievable rate with cost per symbol not exceeding  $\beta$  if for every  $\gamma > 0$  there exists  $n_0$  such that if  $n \geq n_0$ , then an  $(n, M, n\beta, \epsilon)$  code can be found whose rate satisfies  $\log M > n(R - \gamma)$ . Furthermore,  $R$  is said to be achievable if it is  $\epsilon$ -achievable for all  $0 < \epsilon < 1$ . The maximum achievable rate with cost per symbol not exceeding  $\beta$  is the channel capacity denoted by  $C(\beta)$ . The function  $C: \mathbf{R}^+ \rightarrow \mathbf{R}^+$  is referred to as the capacity-cost function [15].

In this paper we restrict attention to memoryless stationary channels. The following is well known [9].

*Theorem 1:* The capacity-cost function of an input-constrained memoryless stationary channel is equal to

$$C(\beta) = \sup_{\substack{X \\ E[b[X]] \leq \beta}} I(X;Y)$$

where the supremum is taken to be zero if the set of distributions therein is empty.

As we saw in the introduction, in addition to (or in lieu of) the maximum number of bits per symbol that can be transmitted with average cost per symbol not exceeding  $\beta$ , it is of interest to investigate the maximum number of bits per unit cost that can be transmitted through the channel. The inverse of this quantity gives the minimum cost of sending one bit of information regardless of how many degrees of freedom it takes to do so. We introduce the following formal definition of this fundamental limit of information transmission.

**Definition 2:** Given  $0 \leq \epsilon < 1$ , a nonnegative number<sup>1</sup>  $R$  is an  $\epsilon$ -achievable rate per unit cost if for every  $\gamma > 0$ , there exists  $\nu_0 \in R^+$  such that if  $\nu \geq \nu_0$ , then an  $(n, M, \nu, \epsilon)$  code can be found with  $\log M > \nu(R - \gamma)$ . Similarly to Definition 1,  $R$  is achievable per unit cost if it is  $\epsilon$ -achievable per unit cost for all  $0 < \epsilon < 1$ , and the capacity per cost is the maximum achievable rate per unit cost.

Note that Definition 2 places no direct penalty or constraint on the number of symbols (or degrees of freedom) used by the code. However, we will see in the corollary to Theorem 2 that the channel capacity per unit cost is not reduced if the blocklength is constrained to grow linearly with the cost, and, thus, logarithmically with the number of messages. Our first result represents the capacity per unit cost in terms of the capacity-cost function.

**Theorem 2:** The capacity per unit cost of a memoryless stationary channel is equal to

$$C = \sup_{\beta > 0} \frac{C(\beta)}{\beta} = \sup_X \frac{I(X; Y)}{E[b[X]]}.$$

*Proof:* First, we show that for every  $\beta > 0$ ,  $C(\beta)/\beta$  is an achievable rate per unit cost. Fix  $\gamma > 0$ ; since  $C(\beta)$  is  $\epsilon$ -achievable there exists  $n_\beta$  such that if  $n > n_\beta$ , then an  $(n, M, n\beta, \epsilon)$  code can be found with

$$\frac{\log M}{n} > C(\beta) - \frac{\gamma\beta}{2}.$$

Now, Let  $\nu_0 = \beta \max\{n_\beta, 2C(\beta)/\gamma\beta\}$ . If  $\nu = n\beta$  and  $\nu \geq \nu_0$ , then that code satisfies

$$\frac{\log M}{\nu} > \frac{C(\beta)}{\beta} - \frac{\gamma}{2};$$

if  $n\beta < \nu < n\beta + \beta$ , then the same code is an  $(n, M, \nu, \epsilon)$  code with

$$\begin{aligned} \frac{\log M}{\nu} &= \frac{\log M}{n\beta} \cdot \frac{n\beta}{\nu} \\ &> \left[ \frac{C(\beta)}{\beta} - \frac{\gamma}{2} \right] \left/ \left[ 1 + \frac{\gamma\beta}{2C(\beta)} \right] \right. \\ &> \frac{C(\beta)}{\beta} - \gamma. \end{aligned} \quad (3)$$

<sup>1</sup>Information per symbol and information per unit cost are differentiated by lightface and boldface characters, respectively.

This shows that  $C(\beta)/\beta$  is  $\epsilon$ -achievable per unit cost for every  $0 < \epsilon < 1$ . Since the set of achievable rates per unit cost is closed,  $\sup_{\beta > 0} C(\beta)/\beta$  is an achievable rate per unit cost.

To prove the converse part, we use the Fano inequality, which implies that every  $(n, M, \nu, \epsilon)$  code satisfies

$$(1 - \epsilon) \log M \leq I(\tilde{X}^n; \tilde{Y}^n) + \log 2 \quad (4)$$

where  $\tilde{X}^n$  is the distribution of the  $n$  input symbols when the messages are equiprobable. Because of the constraint on the cost of each codeword,  $\tilde{X}^n$  satisfies  $E[\sum_{i=1}^n b[\tilde{X}_i]] \leq \nu$ , and thus (4) implies

$$\begin{aligned} \frac{\log M}{\nu} &\leq \frac{1}{1 - \epsilon} \left\{ \frac{n}{\nu} \sup_{\substack{X^n \\ E[\frac{1}{n}\sum_{i=1}^n b[X_i]] \leq \frac{\nu}{n}}} \frac{1}{n} I(X^n; Y^n) + \frac{\log 2}{\nu} \right\} \\ &\leq \frac{1}{1 - \epsilon} \left\{ \frac{n}{\nu} \sup_{\substack{X \\ E[b[X]] \leq \frac{\nu}{n}}} I(X; Y) + \frac{\log 2}{\nu} \right\} \\ &\leq \frac{1}{1 - \epsilon} \left\{ \sup_{\beta > 0} \frac{1}{\beta} \sup_{\substack{X \\ E[b[X]] \leq \beta}} I(X; Y) + \frac{\log 2}{\nu} \right\} \end{aligned}$$

where the intermediate inequality follows from the memorylessness of the channel in the usual way. This implies that if  $R$  is  $\epsilon$ -achievable per unit cost, then for every  $\gamma > 0$ , there exists  $\nu_0$  such that if  $\nu > \nu_0$  then

$$R - \gamma < \frac{1}{1 - \epsilon} \left\{ \sup_{\beta > 0} \frac{C(\beta)}{\beta} + \frac{\log 2}{\nu} \right\}$$

or in other words, for every  $\gamma > 0$ ,

$$R - \gamma \leq \liminf_{\nu \rightarrow \infty} \frac{1}{1 - \epsilon} \left\{ \sup_{\beta > 0} \frac{C(\beta)}{\beta} + \frac{\log 2}{\nu} \right\}.$$

Therefore, if  $R$  is achievable per unit cost, then

$$R \leq \sup_{\beta > 0} \frac{C(\beta)}{\beta}$$

and the theorem follows.  $\square$

**Corollary:** Rate  $R$  is achievable per unit cost if and only if for every  $0 < \epsilon < 1$  and  $\gamma > 0$ , there exist  $s > 0$  and  $\nu_0$ , such that if  $\nu \geq \nu_0$ , then an  $(n, M, \nu, \epsilon)$  code can be found with  $\log M > \nu(R - \gamma)$  and  $n < s\nu$ .

*Proof:* The condition is stronger than that in Definition 2; therefore, according to Theorem 2, it suffices to show that the condition is satisfied for

$$R = \sup_{\beta > 0} \frac{C(\beta)}{\beta}$$

if this quantity is finite, and for arbitrary positive  $R$  otherwise. Fix arbitrary  $0 < \epsilon < 1$  and  $\gamma > 0$ , find  $\beta_0 > 0$

such that  $C(\beta_0)/\beta_0 > R - \gamma$ , and use the code found in the proof of the direct part of Theorem 2 particularized to  $\beta = \beta_0$ . That code is an  $(n, M, \nu, \epsilon)$  code, which satisfies

$$\frac{\log M}{\nu} > \frac{C(\beta_0)}{\beta_0} - \gamma > R - 2\gamma$$

and

$$n \leq \frac{\nu}{\beta_0}. \quad \square$$

It is tempting to conjecture that the restriction to memoryless channels is superfluous for the validity of  $C = \sup C(\beta)/\beta$ . However, even though this formula indeed holds for many channels with memory, it is not universally valid. For example, consider a binary channel with cost function  $b[0] = 0, b[1] = 1$ , and such that

$$y_i = x_i \text{ OR } x_{i-1} \text{ OR } x_{i-2} \text{ OR } \dots$$

The set of all codewords of blocklength  $n$  that contain only one nonzero bit results in an  $(n, n, 1, 0)$  code (and, thus, in an  $(n, n, \nu, \epsilon)$  code for  $\nu \geq 1, 0 \leq \epsilon < 1$ ). Therefore, the capacity per unit cost is infinite (cf. Definition 2) whereas the capacity-cost function is identically zero because there are only  $(n + 1)$  distinct output codewords of blocklength  $n$ .

A generalization of the Arimoto-Blahut algorithm [2] has been given in [12] for the computation of  $\sup_X I(X; Y)/E[b[X]]$  in the case when the input and output alphabets are finite sets and  $b[x] > 0$  for every  $x \in A$ . The motivation for computing this quantity in [12] arose from [19] and [16], which considered channels where the symbol durations need not coincide; hence, the positive constraint on the cost function (equal to the symbol duration in this case).

The special case  $b[x] = 1, x \in A$  renders the capacity per unit cost equal to the conventional capacity of the unconstrained channel. So unless some added structure is imposed, our formulation remains a more general problem than finding the conventional capacity and little more can be said beyond Theorem 2.

It turns out that the additional structure that makes the problem amenable to further interesting analysis is the case when there is a free input symbol, i.e.,  $b[x_0] = 0$  for some  $x_0 \in A$ . For notational convenience, we label the free symbol by 0, and we denote  $A' = A - \{0\}$ . (If there are several free symbols, it is immaterial which one is so labeled.) In this case, it is no longer necessary to find the capacity-cost function  $C(\beta)$  in order to find the capacity per unit cost, and in fact, the problem of computing the capacity per unit cost is usually far easier than that of computing  $C(\beta)$ . Our main result is the following.

**Theorem 3:** If there is a free input symbol, i.e.,  $b[0] = 0$ , then the capacity per unit cost of a memoryless stationary channel is equal to

$$C = \sup_{x \in A'} \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b[x]} \quad (5)$$

where  $P_{Y|X=x}$  denotes the output distribution given that the input is  $x$ , and  $D(P \| Q)$  is the divergence between two measures defined as

$$D(P \| Q) = \begin{cases} \int \log \left( \frac{dP}{dQ} \right) dP, & \text{if } P \ll Q \\ +\infty, & \text{otherwise.} \end{cases} \quad (6)$$

*Proof:* The capacity-cost function  $C(\beta)$  is concave on  $(\beta_{\min}, +\infty)$  where  $\beta_{\min} = \inf b[x]$  [15]. But  $\beta_{\min} = 0$  because there is a free input symbol, and therefore, the function  $C(\beta)/\beta$  is monotone nonincreasing in  $(0, +\infty)$  and

$$C = \sup_{\beta > 0} \frac{C(\beta)}{\beta} = \lim_{\beta \downarrow 0} \frac{C(\beta)}{\beta}. \quad (7)$$

First, consider the case when the free symbol is not unique, i.e., there exists  $x_0 \in A$  such that  $b[x_0] = 0$  and  $D(P_{Y|X=x_0} \| P_{Y|X=0}) > 0$  (if  $P_{Y|X=x_0} = P_{Y|X=0}$ , then  $x_0$  is equivalent to 0 for all purposes and it is excluded from  $A'$ ). Then, any distribution  $X$  such that  $0 < P[X=0] = 1 - P[X=x_0] < 1$  achieves  $I(X; Y) > 0$  and  $E[b[X]] = 0$ . Consequently,  $C(\beta)$  is bounded away from zero on  $[0, +\infty)$  and the right sides of both (7) and (5) are equal to  $+\infty$ .

In the case when the free symbol is unique, we will lower bound  $C(\beta)$  by computing the mutual information achieved by the input distribution  $X$  such that

$$P[X=x_0] = \frac{\beta}{b[x_0]} = 1 - P[X=0] \quad (8)$$

for an arbitrary  $x_0 \in A$ , with  $b[x_0] > 0$ .

If  $D(P_{Y|X=x_0} \| P_{Y|X=0}) = +\infty$ , then

$$\frac{I(X; Y)}{\beta} \geq \frac{D(P_{Y|X=x_0} \| P_Y)}{b[x_0]} \quad (9)$$

and by the continuity of divergence [18, p. 20] the right side of (9) grows without bound as  $\beta \rightarrow 0$  (and  $P_Y \rightarrow P_{Y|X=0}$ ).

If  $D(P_{Y|X=x_0} \| P_{Y|X=0}) < +\infty$ , then we use the general representation

$$\begin{aligned} I(X; Y) &= \int D(P_{Y|X=x} \| P_Y) dP_X(x) \\ &= \int \int \left[ \log \frac{dP_{Y|X=x}(y)}{dP_{Y|X=0}(y)} \right. \\ &\quad \left. + \log \frac{dP_{Y|X=0}(y)}{dP_Y(y)} \right] dP_{Y|X=x}(y) dP_X(x) \\ &= \int D(P_{Y|X=x} \| P_{Y|X=0}) dP_X(x) - D(P_Y \| P_{Y|X=0}), \end{aligned} \quad (10)$$

which is meaningful when the second term in the right side is finite. Particularizing (10) to the choice of input

distribution in (8) we obtain

$$\frac{1}{\beta} I(X; Y) = \frac{1}{b[x_0]} D(P_{Y|X=x_0}) - \frac{1}{\beta} D(P_Y \| P_{Y|X=0}). \tag{11}$$

The limit as  $\beta \rightarrow 0$  of the second term in the right side of (11) is given by the following auxiliary result.

*Lemma:* For any pair of probability measures such that  $P_1 \ll P_0$ , the directional derivative

$$\lim_{\beta \downarrow 0} \frac{1}{\beta} D(\beta P_1 + (1 - \beta) P_0 \| P_0) = 0.$$

*Proof of Lemma:* This result is a special case of identity (2.13) in [5]. We give a self-contained proof for the sake of completeness. For notational convenience, let us introduce the dominating measure  $\mu$  such that  $P_0 \ll \mu, P_1 \ll \mu$  (e.g.,  $\mu = P_0 + P_1$ ), and denote

$$p_k = \frac{dP_k}{d\mu}, \quad k = 0, 1$$

$p_\beta = \beta p_1 + (1 - \beta) p_0$  and  $g_\beta = \frac{1}{\beta} [p_\beta \log(p_\beta / p_0)]$ . Then

$$\frac{1}{\beta} D(\beta P_1 + (1 - \beta) P_0 \| P_0) = \int g_\beta(x) d\mu(x). \tag{12}$$

but

$$\begin{aligned} g_\beta &= \frac{1}{\beta} [p_\beta \log p_\beta - p_0 \log p_0 + p_0 \log p_0 - p_\beta \log p_0] \\ &= \frac{1}{\beta} [p_\beta \log p_\beta - p_0 \log p_0] + (p_0 - p_1) \log p_0. \end{aligned} \tag{13}$$

Since  $p_\beta \log p_\beta$  is convex in  $\beta$ , the chord lemma (e.g., [20, p. 108]) implies that the first term in the right side of (13) is monotone nondecreasing in  $\beta$ . Its limit as  $\beta \downarrow 0$  is equal to

$$\frac{\partial}{\partial \beta} p_\beta \log p_\beta |_{\beta=0} = (p_1 - p_0) \log(e p_0),$$

which together with (13) implies that

$$\lim_{\beta \downarrow 0} g_\beta = (p_1 - p_0) \log e$$

where the convergence is monotone nonincreasing. Therefore the monotone convergence theorem gives the desired result upon substitution in (12).  $\square$

The conclusion from (11) and the Lemma is that

$$C = \lim_{\beta \downarrow 0} \frac{C(\beta)}{\beta} \geq \lim_{\beta \downarrow 0} \frac{I(X; Y)}{\beta} = \frac{D(P_{Y|X=x_0} \| P_{Y|X=0})}{b[x_0]}$$

and consequently,

$$C \geq \sup_{x \in A'} \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b[x]}.$$

To prove the reverse inequality, first consider the upper

bound

$$I(X; Y) \leq \int D(P_{Y|X=x} \| P_{Y|X=0}) dP_X(x); \tag{14}$$

if  $D(P_Y \| P_{Y|X=0}) < \infty$ , then (14) follows from (10) and the nonnegativity of divergence; otherwise the right side of (14) is equal to  $+\infty$  as a result of the convexity of divergence in its first argument. Applying (14) and recalling that in this part of the proof the free symbol is assumed unique, we get that for every  $\beta > 0$

$$\begin{aligned} \frac{C(\beta)}{\beta} &= \frac{1}{\beta} \sup_{\substack{X \\ E[b[X]] \leq \beta}} I(X; Y) \\ &\leq \sup_X \int_{A'} \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b[x]} \frac{b[x]}{\beta} dP_X(x) \\ &\quad \frac{1}{\beta} \int_{b[x] dP_X(x) \leq 1} \\ &\leq \sup_{x \in A'} \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{b[x]} \end{aligned} \tag{15}$$

concluding the proof of Theorem 3.  $\square$

We have seen in the foregoing proof that to transmit information at minimum cost it is enough to restrict attention to codes which use only one symbol in the input alphabet in addition to 0. We will highlight this fact with an alternative proof of the achievability of  $\sup D(P_{Y|X=x} \| P_{Y|X=0}) / b[x]$  which unlike that in Theorem 3, does not hinge on the result of a conventional coding theorem (via Theorem 2). This alternative proof of the direct part of our coding theorem does not follow any of the approaches known for proving conventional direct coding theorems, viz., typical sequences, types, random coding exponent and Feinstein's lemma. Rather, it is a direct corollary of Stein's lemma [4] on the asymptotic error probability of binary hypothesis tests with i.i.d. observations. Interestingly, Stein's lemma can be viewed as a generalization of Shannon's *first* (source coding) theorem (cf. [6]).

*Codebook:* Fix  $x_0 \in A'$  and integers  $M > 0$  and  $N > 0$ . Codeword  $m \in \{1, \dots, M\}$  corresponds to an  $M \times N$  matrix all whose columns are identical to  $[r_1, \dots, r_M]^T$  where

$$r_i = \begin{cases} x_0, & i = m \\ 0, & i \neq m. \end{cases}$$

Thus the blocklength of this code is  $n = MN$  and the cost of each codeword is  $\nu = Nb[x_0]$ .

*Decoding:* Fix  $0 < \epsilon < 1$ . The decoder observes the matrix of independent observations  $\{y_{ij}, 1 \leq i \leq M, 1 \leq j \leq N\}$  where  $y_{ij}$  has distribution  $P_{Y|X=x_0}$  if  $i = m$  and  $P_{Y|X=0}$  otherwise, given that codeword  $m$  is transmitted. The decoder is not maximum-likelihood, rather it performs  $M$  independent binary hypothesis tests:

$$H_{i0}: r_i = 0$$

$$H_{i1}: r_i = x_0$$

$i = 1, \dots, M$ . The conditional error probabilities of those

tests are denoted by

$$\begin{aligned}\alpha_{iN} &= P[\hat{r}_i = x_0 | r_i = 0] \\ \beta_{iN} &= P[\hat{r}_i = 0 | r_i = x_0]\end{aligned}$$

and the thresholds are set so that  $\beta_{iN} \leq \frac{\epsilon}{2}$ . Then, message  $m$  is declared if

$$\hat{r}_i = \begin{cases} x_0, & i = m \\ 0, & i \neq m \end{cases}$$

otherwise (i.e., if  $\hat{r}_i \neq 0$  for zero or more than one  $i$ ), an error is declared.

To evaluate the rate per unit cost that can be achieved with this code, we need to estimate how large  $M$  can be so that the average probability of error does not exceed  $\epsilon$ . Obviously, the probability of error conditioned on message  $m$  being sent  $P_m$  is independent of the value of  $m$  and can be upper bounded by

$$P_1 \leq \beta_{1N} + (M-1)\alpha_{1N}.$$

By Stein's lemma (see [2, Theorem 4.4.4]), since  $\beta_{1N} \leq \frac{\epsilon}{2}$ , for every  $\gamma > 0$  we can achieve

$$\alpha_{1N} \leq \exp \left[ -ND(P_{Y|X=x_0} \| P_{Y|X=0}) + N\gamma \right]$$

for all sufficiently large  $N$ . Consequently, if

$$\frac{\log M}{Nb[x_0]} < \frac{D(P_{Y|X=x_0} \| P_{Y|X=0})}{b[x_0]} - \frac{2\gamma}{b[x_0]}$$

then,  $P_1 \leq \epsilon$  for sufficiently large  $N$ , and therefore,<sup>2</sup>  $D(P_{Y|X=x_0} \| P_{Y|X=0})/b[x_0]$  is an achievable rate per unit cost.

Notice that the dependence of the blocklength of the foregoing code in the number of messages is superlinear:  $O(M \log M)$ ; however, the corollary to Theorem 2 guarantees that even if the blocklength is constrained to grow logarithmically in the number of messages, the same rate per unit cost can be achieved.

The preceding codebook is orthogonal in the sense that the nonzero components of different codewords do not overlap (cf. [10]). This corresponds to pulse position modulation (PPM) either in the time or frequency domains. Interestingly, the realization that PPM can achieve the maximum number of bits per unit energy in the Gaussian channel goes back to Golay [11], who used a code essentially equivalent to the foregoing.

The quantity

$$\sup_{\substack{x \in A' \\ c[x] \leq \beta}} D(P_{Y|X=x} \| P_{Y|X=0})$$

can be interpreted as the maximum ratio of transmitted bits to (nonzero) degrees of freedom used per codeword subject to a (peak) constraint on the cost of each symbol dictated by a cost function  $c: A' \rightarrow R^+$ . To see this, particularize the result of Theorem 3 to the case:  $b[x] = 1$  if  $x \neq 0$  and input alphabet equal to  $\{x \in A, c[x] \leq \beta\}$ .

<sup>2</sup>This only shows existence of the desired codes with cost-per-codeword  $\nu$  equal to a multiple of  $b[x_0]$ . Intermediate values of  $\nu$  are handled as in (3).

As we have mentioned, one of the nice features of the capacity per unit cost is that it is easier to compute than the ordinary capacity, for channels with a free input symbol. We now give several illustrative examples.

*Example 1 (Gallager's energy limited change):* In [10], Gallager considers a discrete-time stationary memoryless channel with input alphabet  $\{0, 1\}$  and real-valued output with density function  $p_i$  conditioned on the input being equal to  $i = 0, 1$ . The energy function is  $b[0] = 0$  and  $b[1] = 1$ . Using Theorem 3, we get that the capacity per unit cost is equal to

$$C = D(p_1 \| p_0) = \int_{-\infty}^{\infty} p_1(x) \log \frac{p_1(x)}{p_0(x)} dx.$$

The reliability function of the rate per unit cost  $\tilde{E}(R)$  of this channel is found exactly in [10], where it is shown that  $\tilde{E}(R) > 0$  if and only if  $R$  is smaller than a certain parameter  $\tilde{E}'_0(0)$  defined therein. It can be checked that  $\tilde{E}'_0(0)$  coincides with the divergence between  $p_1$  and  $p_0$ .

The case of a finite (nonbinary) input alphabet  $\{0, \dots, K\}$  with arbitrary energy function for nonzero inputs is also considered in [10]. In this case, the reliability function of the rate per unit cost is only known exactly under a certain sufficient condition. Nevertheless, as in the binary case, the random-coding lower bound and the sphere-packing upper bound to the reliability function are equal to zero for rates per unit cost greater than  $\tilde{E}'_0(0)$  that can be shown to coincide with  $\max_{1 \leq j \leq K} D(p_j \| p_0)/b[j]$ .

*Example 2 (Poisson Channel):* In the continuous-time Poisson channel studied in [7], [22], the channel inputs are functions  $\{\lambda(t), t \in [0, T]\}$  such that  $0 \leq \lambda(t) \leq \lambda_1$  and the channel output is a Poisson point process with rate  $\lambda(t) + \lambda_0$ . The cost of the input waveform  $\{\lambda(t), t \in [0, T]\}$  is equal to  $\int_0^T \lambda(t) dt$ , which can be interpreted as the average number of photons detected by a photodetector when the instantaneous magnitude squared of the electric field incident on its screen is  $\lambda(t)$ . Therefore, in this case, the capacity per unit cost can be interpreted as the maximum number of bits that can be transmitted per photon. To put this channel in an equivalent discrete-time formulation which fits in our framework, fix  $T_0 > 0$ , consider the input alphabet to be the set of waveforms  $\{x(t), t \in [0, T_0], 0 \leq x(t) \leq \lambda_1\}$ , and let  $b[\{x(t)\}] = \int_0^{T_0} x(t) dt$ . Hence there is a free symbol:  $\{x(t) = 0, t \in [0, T_0]\}$ . Since we place no restrictions on the variation of each of these waveforms on  $[0, T_0]$ , it is clear that if we let  $T = nT_0$  and let a codeword  $(\{x_1(t), \dots, x_n(t)\})$  correspond with the continuous-time input  $\lambda(t) = \sum_{i=0}^{n-1} x_i(t - iT_0)$  the model is equivalent to the original one (the fact that  $T$  grows as a multiple of  $T_0$  can be easily overcome by appending a zero-valued waveform of duration smaller than  $T_0$ ).

The divergence between the outputs due to input waveforms  $\{x(t), t \in [0, T_0]\}$  and  $\{0, t \in [0, T_0]\}$  can be computed using the sample function density of a Poisson point process with rate  $\{\lambda(t), t \in [0, T_0]\}$  with respect to

the probability measure generated by a unit-rate Poisson point process evaluated at the unordered realization  $(t_1, \dots, t_K)$  [3]:

$$p_\lambda(t_1, \dots, t_K) = \exp_e \left( \int_0^{T_0} [1 - \lambda(t)] dt \right) \prod_{i=1}^K \lambda(t_i).$$

Then

$$D(p_x \| p_0) = E \left[ \log \frac{\exp_e \left( - \int_0^{T_0} [x(t) + \lambda_0] dt \right) \prod_{i=1}^K [x(t_i) + \lambda_0]}{\exp_e \left( - \int_0^{T_0} \lambda_0 dt \right) \prod_{i=1}^K \lambda_0} \right]$$

$$= - \int_0^{T_0} x(t) dt \log e + E \left[ \sum_{i=1}^K \log \left( 1 + \frac{x(t_i)}{\lambda_0} \right) \right]$$

where the expectation is with respect to the measure generated by the Poisson process with rate  $\{x(t) + \lambda_0, t \in [0, T_0]\}$ . A straightforward computation shows

$$D(p_x \| p_0) = - \int_0^{T_0} x(t) dt \log e + \int_0^{T_0} (x(t) + \lambda_0) \log \left( 1 + \frac{x(t)}{\lambda_0} \right) dt$$

and by virtue of Theorem 3 the capacity per unit cost is

$$C = \sup_{\substack{0 \leq x(t) \leq \lambda_1 \\ 0 \leq t \leq T_0}} \frac{\int_0^{T_0} (x(t) + \lambda_0) \log \left( 1 + \frac{x(t)}{\lambda_0} \right) dt}{\int_0^{T_0} x(t) dt} - \log e$$

$$= \max_{0 \leq z \leq \lambda_1} \left( 1 + \frac{\lambda_0}{z} \right) \log \left( 1 + \frac{z}{\lambda_0} \right) - \log e$$

$$= \left( 1 + \frac{\lambda_0}{\lambda_1} \right) \log \left( 1 + \frac{\lambda_1}{\lambda_0} \right) - \log e,$$

which can be shown to coincide with

$$\sup_{\rho > 0} \frac{C(\rho, \lambda_1)}{\rho \lambda_1} = \lim_{\rho \downarrow 0} \frac{C(\rho, \lambda_1)}{\rho \lambda_1}$$

where  $C(\rho, \lambda_1)$  is the capacity of the channel with average cost per codeword not exceeding  $\rho \lambda_1$  found in [7], [22].

*Example 3 (Gaussian Channel):* Consider the discrete-time memoryless channel with additive and multiplicative Gaussian noise

$$y_i = (\alpha + z_i) x_i + w_i$$

where the input symbol is  $x_i \in R$  and  $\{z_i\}$  and  $\{w_i\}$  are independent i.i.d. Gaussian sequences with zero-mean and variances  $\gamma^2$  and  $\sigma^2$ , respectively. The conventional additive white noise channel corresponds to the case  $\gamma = 0$ ; whereas the case  $\alpha = 0$  arises (in a multidimensional version) in the modeling of fading dispersive channels [9]. The divergence between two Gaussian distribu-

tion is easily shown to be given by

$$D(N(m_1, \sigma_1^2) \| N(m_0, \sigma_0^2)) = \log \frac{\sigma_0}{\sigma_1} + \left[ \frac{\sigma_1^2}{2\sigma_0^2} + \frac{(m_1 - m_0)^2}{2\sigma_0^2} - \frac{1}{2} \right] \log e,$$

which implies

$$D(P_{Y|X=x} \| P_{Y|X=0})$$

Hence, if  $b[x] = x^2$ , then the capacity per unit cost is

$$C = \frac{1}{2} \frac{\gamma^2 + \alpha^2}{\sigma^2} \log e.$$

At this point it may be worthwhile stating a simple lower bound to the capacity per unit cost in the special case when  $A = R$  and  $b[x] = x^2$ . Within mild regularity conditions on the family of conditional distributions  $\{P_{Y|X=x}; x \in R\}$ , the following asymptotic result on the divergence is known [14, 2.6]

$$\lim_{x \downarrow 0} \frac{D(P_{Y|X=x} \| P_{Y|X=0})}{x^2} = \frac{1}{2} I_0(P_{Y|X})$$

where  $I_x(P_{Y|X})$  is Fisher's information for estimating  $X$  from  $Y$ . From Theorem 3, it follows that<sup>3</sup>

$$C \geq \frac{1}{2} I_0(P_{Y|X}). \tag{16}$$

Coupling (16) to the Cramer-Rao bound [2, 8.1.3] we obtain an interesting connection between information theory and estimation theory: the minimum energy necessary to transmit one half nat information (0.721 bits) through the channel cannot exceed the minimum conditional variance of an estimate of the input from the output given that the input is 0.

The reason for this connection can be explained as follows. The capacity (and, hence, the capacity per unit cost) of the channel cannot increase if the decoder is constrained to use, in lieu of the channel outputs, the input estimates provided on a symbol by symbol basis by an unbiased minimum variance estimator. Then, the channel seen by the decoder (Fig. 2) can be viewed as a memoryless channel with additive noise (possibly dependent of the input), i.e., the output of the equivalent

<sup>3</sup>If  $A = R^K$  and  $b[x] = \|x\|^2$ , then  $I_0(P_{Y|X})$  is replaced by the largest eigenvalue of Fisher's information matrix.

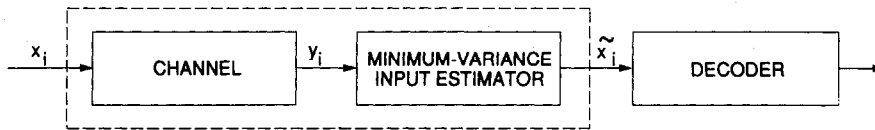


Fig. 2. Modified decoder for justification of estimation-theoretic bound.

channel can be written as  $\tilde{X}_i = X_i + N_i$ , with  $\text{var}(N_i) = \sigma^2$ . The capacity of this channel cannot be smaller than the capacity of an additive white Gaussian noise channel with noise variance equal to  $\sigma^2$ . This can be checked by generalizing Theorem 7.4.3 in [9] to the case where the additive noise and the input are uncorrelated, but not necessarily independent. Therefore, the minimum energy necessary to send one bit through the channel cannot be greater than  $\sigma^2 \ln 4 = \sigma^2 / 0.721$ . But for reliable communication with minimum energy, the overwhelming majority of inputs are equal to 0 and  $\sigma^2$  is arbitrarily close to the minimum conditional estimate of the input from the output given that the input is 0.

Note that this estimation-theoretic bound on the minimum energy required to transmit information reliably is satisfied with equality in the case of the additive Gaussian channel because in that case the Cramer-Rao bound is tight and the input estimator gives a scaled version of the channel output.

Another simple bound to the capacity per unit cost is obtained using Kullback's inequality [13],

$$D(P_{Y|X=x} \| P_{Y|X=0}) \geq \frac{1}{2} \left( \int |P_{Y|X=x} - P_{Y|X=0}| \right)^2.$$

The lower bound in (16) is satisfied with equality in the additive Gaussian noise channel (Example 3) and in the next example which illustrates the simplicity of the computation of capacity per unit cost in channels where the computation of the capacity-cost function is a pretty hopeless task.

*Example 4 (Additive non-Gaussian noise channel):* Consider the discrete-time memoryless channel with additive Cauchy noise  $y_i = x_i + n_i$  where the probability density function of each noise sample is

$$f(z) = \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{z}{\sigma}\right)^2}.$$

The divergence between two shifted versions of the Cauchy distribution is easily computed

$$\begin{aligned} D(P_{Y|X=x} \| P_{Y|X=0}) &= \int_{-\infty}^{\infty} f(z-x) \log \frac{f(z-x)}{f(z)} dz \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+t^2} \log \frac{1 + \left(t + \frac{x}{\sigma}\right)^2}{1+t^2} dt \\ &= \log \left( 1 + \frac{x^2}{4\sigma^2} \right), \end{aligned}$$

which results in

$$C = \frac{1}{4\sigma^2} \log e,$$

when the cost is equal to the energy ( $b[x] = x^2$ ).

A similar exercise for Laplace distributed noise with probability density function

$$f(z) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}|z|/\sigma}$$

yields

$$D(P_{Y|X=x} \| P_{Y|X=0}) = \left( \frac{\sqrt{2}}{\sigma} |x| - 1 + e^{-\sqrt{2}|x|/\sigma} \right) \log e$$

and

$$C = \frac{1}{\sigma^2} \log e,$$

which satisfies (16) with equality as well.<sup>4</sup>

### III. CAPACITY REGION PER UNIT COST OF MULTIUSER CHANNELS

We consider frame-synchronous discrete-time memoryless two-user multiple-access channels with arbitrary input alphabets  $A_1$  and  $A_2$  and output alphabet  $B$ . (At the end of this section we generalize the results obtained for the multiple-access channel assuming the existence of free symbols to the interference channel.) A  $(n, M_1, M_2, \nu_1, \nu_2, \epsilon)$  code has blocklength  $n$ ; average probability of decoding both messages correctly better than  $1 - \epsilon$ ;  $M_k$  codewords for user  $k$ ; and each codeword of user  $k$  satisfies  $\sum_{i=1}^n b_k[x_{mi}] \leq \nu_k$  where  $b_k: A_k \rightarrow R^+$ , and  $k = 1, 2$ . Generalizing Definition 1 to the multiple-access channel (cf. [6]) and denoting the capacity region of the stationary memoryless two user channel with cost per symbol for user  $k$  not exceeding  $\beta_k$ ,  $k = 1, 2$ , by  $C(\beta_1, \beta_2)$ , the following coding theorem holds.

*Theorem 4 [10]:* Define the directly achievable region,

$$\begin{aligned} A(\mu_1, \mu_2) &= \bigcup_{\substack{X_1, X_2 \\ E[b_k[X_k]] \leq \mu_k \\ k=1,2}} \{ (R_1, R_2) : 0 \leq R_1 \leq I(X_1; Y|X_2) \\ &\quad 0 \leq R_2 \leq I(X_2; Y|X_1) \\ &\quad R_1 + R_2 \leq I(X_1, X_2; Y) \} \end{aligned} \quad (17)$$

where the union is over independent distributions on the input alphabets.

<sup>4</sup>A sufficient condition for (16) to be satisfied with equality in an energy-constrained additive noise channel is that the noise density be even and its convolution with the fourth derivative of its logarithm be nonnegative. (This condition was obtained jointly with H. V. Poor.)

Then,  $C(\beta_1, \beta_2)$  is equal to the closure of

$$\left\{ (R_1, R_2) : ((R_1, R_2), (\beta_1, \beta_2)) \in \text{convex} \right. \\ \left. \cdot \bigcup_{\substack{\mu_1 > 0 \\ \mu_2 > 0}} (A(\mu_1, \mu_2), (\mu_1, \mu_2)) \right\}. \quad (18)$$

The direct part of this result is a straightforward generalization of the unconstrained version of the result [6], [8]; the converse is due to Gallager [10] along with the realization that  $C(\beta_1, \beta_2)$  need not equal  $A(\beta_1, \beta_2)$  (a longstanding common misconception). It can be shown that (18) can be written as

$$\bigcup_{\substack{X_1, X_2, V \\ E[b_k[X_k]] \leq \beta_k \\ k=1,2}} \left\{ \begin{array}{l} 0 \leq R_1 \leq I(X_1; Y|X_2, V) \\ 0 \leq R_2 \leq I(X_2; Y|X_1, V) \\ R_1 + R_2 \leq I(X_1, X_2; Y|V) \end{array} \right.$$

where  $(X_1, X_2)$  are conditionally independent given the simple random variable  $V$  and  $Y$  is conditionally independent of  $V$  given  $(X_1, X_2)$ . This alternative expression is akin to that given by Csiszár and Körner [6, p. 278] in the unconstrained case.

We define the capacity region per unit cost similarly to Definition 2.

**Definition 3:** Given  $0 < \epsilon < 1$ , a pair of nonnegative numbers  $(R_1, R_2)$  is an  $\epsilon$ -achievable rate pair per unit cost if for every  $\gamma > 0$ , there exists,  $(\bar{\nu}_1, \bar{\nu}_2) \in R^+ \times R^+$  such that if  $x > 1$  then an  $(n, M_1, M_2, \nu_1, \nu_2, \epsilon)$  code can be found with

$$\frac{\log M_k}{\nu_k} \geq R_k - \gamma, \quad k=1,2$$

and  $(\nu_1, \nu_2) = x(\bar{\nu}_1, \bar{\nu}_2)$ . The pair  $(R_1, R_2)$  is achievable per unit cost if it is  $\epsilon$ -achievable per unit cost for all  $0 < \epsilon < 1$  and the capacity region per unit cost is the set of all achievable pairs per unit cost. Theorem 5 next is a generalization of Theorem 2.

**Theorem 5:** The capacity region per unit cost of a memoryless stationary two-user channel is equal to

$$C = \bigcup_{\substack{\beta_1 > 0 \\ \beta_2 > 0}} \{ (R_1, R_2) : (\beta_1 R_1, \beta_2 R_2) \in C(\beta_1, \beta_2) \}. \quad (19)$$

*Proof:* The proof of the direct part is a straightforward generalization of the corresponding proof in Theorem 2. To show the converse part, note that because of the Fano inequality, every  $(n, M_1, M_2, \nu_1, \nu_2, \epsilon)$  code satisfies

$$(1-\epsilon) \left( \frac{\nu_1 \log M_1}{n \nu_1}, \frac{\nu_2 \log M_2}{n \nu_2} \right) \\ \in A_n \left( \frac{\nu_1}{n}, \frac{\nu_2}{n} \right) + \left( \frac{1}{n}, \frac{1}{n} \right) \log 2 \quad (20)$$

where  $A_n$  is the directly achievable region for the  $n$ -block

version of the channel,

$$A_n(\mu_1, \mu_2) = \bigcup_{X_1^n X_2^n} \left\{ \begin{array}{l} 0 \leq R_1 \leq \frac{1}{n} I(X_1^n; Y^n | X_2^n) \\ \frac{1}{n} \sum_{k=1}^n E[b_k[X_{kn}]] \leq \mu_k \\ 0 \leq R_2 \leq \frac{1}{n} I(X_2^n; Y^n | X_1^n) \\ R_1 + R_2 \leq \frac{1}{n} I((X_1^n, X_2^n); Y^n) \end{array} \right\} \quad (21)$$

which is a subset of  $C(\mu_1, \mu_2)$  because the channel is memoryless. From (20) and Definition 3 we have that if  $(R_1, R_2)$  is achievable per unit cost, then for all  $0 < \epsilon < 1$ ,  $\gamma > 0$ ,  $(\nu_1, \nu_2) = x(\bar{\nu}_1, \bar{\nu}_2)$  and  $x > 1$ , there exists  $n$  such that

$$(1-\epsilon) \left( \frac{\nu_1}{n} (R_1 - \gamma), \frac{\nu_2}{n} (R_2 - \gamma) \right) \in C \left( \frac{\nu_1}{n}, \frac{\nu_2}{n} \right) \\ + \left( \frac{1}{\nu_1} \frac{\nu_1}{n}, \frac{1}{\nu_2} \frac{\nu_2}{n} \right) \log 2,$$

which implies

$$(R_1, R_2) \in \bigcup_{\substack{\beta_1 > 0 \\ \beta_2 > 0}} \left\{ (R_1, R_2) : (1-\epsilon)(\beta_1(R_1 - \gamma), \right. \\ \left. \beta_2(R_2 - \gamma)) \in C(\beta_1, \beta_2) + \left( \frac{\beta_1}{\nu_1}, \frac{\beta_2}{\nu_2} \right) \log 2 \right\}$$

and since  $\min\{\nu_1, \nu_2\}$  is arbitrarily large

$$(R_1, R_2) \in \bigcup_{\substack{\beta_1 > 0 \\ \beta_2 > 0}} \{ (R_1, R_2) : (1-\epsilon)(\beta_1(R_1 - 2\gamma), \\ \beta_2(R_2 - 2\gamma)) \in C(\beta_1, \beta_2) \} \quad (22)$$

But since (22) holds for arbitrarily small  $\epsilon$  and  $\gamma$  and the set  $C(\beta_1, \beta_2)$  is closed,  $(R_1, R_2)$  must belong to the right side of (19).  $\square$

Using the same idea as in the corollary to Theorem 2, it is easy to prove that the capacity region per unit cost is not reduced if the blocklength is constrained to grow linearly with the cost.

**Corollary:** The pair  $(R_1, R_2)$  is achievable per unit cost if and only if for every  $0 < \epsilon < 1$  and  $\gamma > 0$ , there exists  $\beta_1 > 0$  and  $\beta_2 > 0$  such that an  $(n, M_1, M_2, n\beta_1, n\beta_2, \epsilon)$  code can be found for all sufficiently large  $n$ , with

$$\frac{\log M_k}{n\beta_k} > R_k - \gamma.$$

As in single-user channels, the interesting case which allows us to proceed beyond Theorem 5 is when each user has a free symbol.

**Theorem 6:** If both alphabets contain free input symbols, i.e.,  $0 \in A_k$  such that  $b_k[0] = 0$ , then the following

rectangle is achievable per unit cost

$$C \supset \left\{ 0 \leq R_1 \leq \sup_{x \in A_1} \frac{D(P_{Y|X_1=x, X_2=0} \| P_{Y|X_1=0, X_2=0})}{b_1[x]} \right\} \\ \times \left\{ 0 \leq R_2 \leq \sup_{x \in A_2} \frac{D(P_{Y|X_1=0, X_2=x} \| P_{Y|X_1=0, X_2=0})}{b_2[x]} \right\}. \quad (23)$$

*Proof:* Consider the two single-user channels that are derived from the multiple-access channel by letting all input symbols from user 2 and user 1, respectively, be equal to 0. Suppose that we have an  $(n, M_k, \nu_k, \epsilon_k)$  code for the  $k$ th single-user channel,  $k=1,2$ . Then we can construct a  $(2n, M_1, M_2, \nu_1, \nu_2, \epsilon_1 + \epsilon_2)$  code for the original multiple-access channel by simply juxtaposing  $n$  0's to the left [resp. right] of each codeword of the single-user code 1 [resp. 2], and by having two independent single-user decoders. Then, the inner bound (23) follows from the single-user result of Theorem 3.  $\square$

Before we find sufficient conditions that guarantee that the inner bound in (23) is equal to the capacity region per unit cost, we exhibit an example where the inclusion in (23) is strict: Let  $A_1 = A_2 = B = \{0, 1\}$ ,  $b_k[0] = 0$ ,  $b_k[1] = 1$ , and  $y_i = x_{1i}$  AND  $x_{2i}$ . The inner bound in Theorem 6 is equal to  $\{(0,0)\}$ , whereas the capacity region per unit cost (computed via Theorem 5) is depicted in Fig. 3. Several observations on this region are in order. First, note that the capacity region per unit cost need not be convex, because the time-sharing principle does not apply here: the rate pair per unit cost of two time-shared codes is not the convex combination of the respective rate pairs per unit cost. Second, note that any rate pair of the form  $(0, R)$  or  $(R, 0)$  is achievable, because if one of the users always sends 1 (an altruistic strategy that incurs in the highest possible cost without sending any information), the other user sees a noiseless binary channel, over which it is possible to send any amount of information with one unit of cost (using sufficiently long blocklength). Third, even though both users have a free symbol, coding strategies that use a very low percentage of nonzeros are not optimum, in contrast to the single-user channel. For ex-

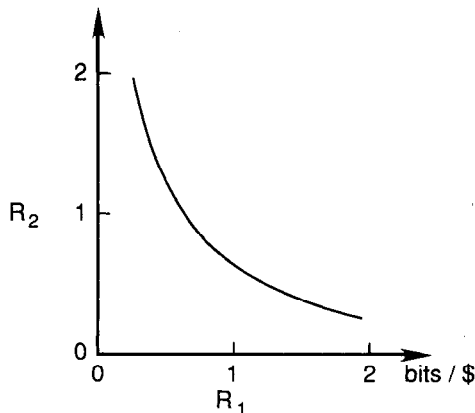


Fig. 3. Capacity region per unit cost of AND multiple-access channel.

ample, the boundary point of the capacity region per unit cost in which both users have the same rate per unit cost, 0.81 bits/\$ (or about 185¢ to send one bit), is achieved with  $P[X_1 = 1] = P[X_2 = 1] = 0.49$ .

The underlying reason why it is not possible to proceed beyond Theorem 5 in the case of the AND channel, is that use of the free symbol by one of the users disables transmission by the other user. However, in many multiple-access channels, the opposite situation is encountered, namely, use of the free symbol by one of the users corresponds to the most favorable single-user channel seen by the other user. In that case, we have the following result.

*Theorem 7:* Let  $C_a^{(1)}(\beta)$  be the capacity-cost function of the single-user channel seen by user 1 when input 2 is equal to the constant  $a \in A_2$ , i.e., a memoryless single-user channel with conditional output probability  $P_{Y|X_1=x, X_2=a}$  (and  $C_a^{(2)}(\beta)$  is defined analogously for user 2).

Suppose that both input alphabets contain a free symbol and that

$$C_0^{(k)}(\beta) = \max_{a \in A_k} C_a^{(k)}(\beta), \quad \text{for } \beta > 0 \text{ and } k = 1, 2. \quad (24)$$

Then the inner bound in Theorem 6 is equal to the capacity region per unit cost.

*Proof:* Let us introduce the set

$$A_0(\mu_1, \mu_2) = [0, C_0^{(1)}(\mu_1)] \times [0, C_0^{(2)}(\mu_2)]. \quad (25)$$

Note that

$$A(\beta_1, \beta_2) \subset A_0(\beta_1, \beta_2) \quad (26)$$

because for any random variables  $X_1, X_2$  such that  $E[b_1[X_1]] \leq \mu_1$

$$I(X_1; Y|X_2) = \int I(X_1; Y|X_2 = a) dP_{X_2}(a) \\ \leq \int C_a^{(1)}(\mu_1) dP_{X_2}(a) \\ \leq C_0^{(1)}(\mu_1).$$

Moreover the concavity of  $C_0^{(k)}(\mu_k)$  ensures that

$$A_0(\beta_1, \beta_2) = \left\{ (R_1, R_2) : ((R_1, R_2), (\beta_1, \beta_2)) \in \text{convex} \right. \\ \left. \bigcup_{\substack{\mu_1 > 0 \\ \mu_2 > 0}} (A_0(\mu_1, \mu_2), (\mu_1, \mu_2)) \right\}$$

and consequently, via (26)

$$C(\beta_1, \beta_2) \subset A_0(\beta_1, \beta_2). \quad (27)$$

Together with Theorem 5 this implies

$$C \subset \bigcup_{\substack{\beta_1 > 0 \\ \beta_2 > 0}} \{(R_1, R_2) : (\beta_1 R_1, \beta_2 R_2) \in A_0(\beta_1, \beta_2)\}$$

and the theorem follows applying the single-user result in Theorem 3.  $\square$

A more stringent sufficient condition for Theorem 7 is that for any input distributions

$$I(X_i; Y | X_j = 0) = \max_{x \in A_j} I(X_i; Y | X_j = x)$$

for  $(i, j) = (1, 2), (2, 1)$ . An important case where this condition is satisfied is the additive channel  $y_i = x_{1i} + x_{2i} + n_i$ . In this case,  $I(X_i; Y | X_j = a)$  is independent of  $a \in A_j$ ,  $A(\beta_1, \beta_2) = C(\beta_1, \beta_2)$ , and the computation of the capacity region per unit cost boils down to the computation of single-user capacity per unit cost.

The decoupled nature of the results in Theorems 6 and 7 makes them an easy target for generalization to the *interference channel*. Theorem 6 and its proof hold verbatim with the rectangle in (23) replaced by

$$C \supset \left\{ 0 \leq R_1 \leq \sup_{x \in A_1} \frac{D(P_{Y_1|X_1=x, X_2=0} \| P_{Y_1|X_1=0, X_2=0})}{b_1[x]} \right\} \\ \times \left\{ 0 \leq R_2 \leq \sup_{x \in A_2} \frac{D(P_{Y_2|X_1=0, X_2=x} \| P_{Y_2|X_1=0, X_2=0})}{b_2[x]} \right\}.$$

Regarding the generalization of Theorem 7 to the interference channel, it is possible to bypass Theorem 5 using the following bounding argument. If both encoder 1 and decoder 1 were informed of the codeword to be sent by encoder 2 prior to transmission, they would see a single-user arbitrarily varying channel with both encoder and decoder informed of the state sequence. The capacity-cost of this channel is equal to [6, p. 227]

$$\inf_{a \in A_2^*} C_a^{(1)}(\beta)$$

for some  $A_2^* \subset A_2$ . Because of (24), this is further upper bounded by  $C_0^{(1)}(\beta)$ . It follows that under this hypothetical situation with side information, the capacity per unit cost achievable by the first user is upper bounded by

$$\sup_{\beta > 0} \frac{C_0^{(1)}(\beta)}{\beta} = \sup_{x \in A_1} \frac{D(P_{Y_1|X_1=x, X_2=0} \| P_{Y_1|X_1=0, X_2=0})}{b_1[x]}.$$

#### IV. RELATED PROBLEMS

A problem that should be examined is whether there is a counterpart to our channel coding result in rate-distortion theory. The rate-distortion theorem [21] states that the minimum number of bits that need to be transmitted per source symbol so as to reproduce the source with average distortion not exceeding  $\delta$  is the rate-distortion function,

$$R(\delta) = \inf_{\substack{P_{Y|X} \\ E[d(X, Y)] \leq \delta}} I(X; Y) \quad (28)$$

where the nonnegative function  $d(x, y)$  assigns a penalty to each input-output pair. Let  $\delta_{\max}$  be such that  $R(\delta_{\max}) = 0$  and  $R(\delta) > 0$  for  $\delta < \delta_{\max}$ , i.e.,  $\delta_{\max}$  is the minimum distortion that can be achieved by representing the source by a fixed symbol;

$$\delta_{\max} = E[d(X, v_X)] \quad (29)$$

with

$$v_X = \arg \min_y E[d(X, y)]. \quad (30)$$

In some situations it may be of interest to find the level of distortion reduction from  $\delta_{\max}$  that can be achieved by low-rate coding. If we consider the reward function  $\delta_{\max} - d(x, y)$ , then the minimum number of bits necessary to get one reward unit is the slope of the rate-distortion function at  $\delta_{\max}$ . The counterpart to Theorem 3 in rate-distortion theory is

$$R'(\delta_{\max}) = \lim_{\delta \uparrow \delta_{\max}} \frac{R(\delta)}{\delta_{\max} - \delta} \\ = \inf_{P_W \ll P_X} \frac{D(P_W \| P_X)}{E[d(W, v_X) - d(W, v_W)]}. \quad (31)$$

The direct part of (31) can be shown by using the following low-rate coding scheme: fix an arbitrary source distribution  $P_W$ ; for every string of  $n$  source symbols the encoder informs the decoder that each symbol should be represented by  $v_W$  if the string is  $P_W$ -typical or by  $v_X$ , otherwise. Since the true distribution is  $P_X$ , the probability that the string is  $P_W$ -typical is (cf. [6, Lemma 2.6]) essentially

$$p = \exp(-nD(P_W \| P_X)). \quad (32)$$

Therefore, it suffices to transmit  $h(p) = -p \log p - (1-p) \log(1-p)$  information units per  $n$  source symbols. On the other hand, (as  $n \rightarrow \infty$ ) the average distortion remains equal to  $\delta_{\max}$  if the string is not  $P_W$ -typical, and it decreases from  $E[d(W, v_X)]$  to  $E[d(W, v_W)]$  when the string is  $P_W$ -typical. Thus, the ratio of rate to distortion reduction is (via (32))

$$\lim_{n \rightarrow \infty} \frac{h(p)}{npE[d(W, v_X) - d(W, v_W)]} \\ = \frac{D(P_W \| P_X)}{E[d(W, v_X) - d(W, v_W)]}.$$

To obtain the converse of (31), note that by the symmetry of mutual information and the first equality in (10)

$$I(X; Y) = \int D(P_{X|Y=y} \| P_X) dP_Y(y).$$

Thus

$$\begin{aligned}
 R'(\delta_{\max}) &\geq \inf_{\substack{P_{X|Y} \ll P_X \\ P_Y}} \frac{\int D(P_{X|Y=y} \| P_X) dP_Y(y)}{\int d(x, v_X) dP_X(x) - \iint P_{X|Y=y}(x) d(x, y) dP_Y(y)} \\
 &\geq \inf_{\substack{P_{X|Y} \ll P_X \\ P_Y}} \frac{\int D(P_{X|Y=y} \| P_X) dP_Y(y)}{\iint P_{X|Y=y}(x) [d(x, v_X) - d(x, v_{X|Y=y})] dP_Y(y)} \\
 &= \inf_{P_W \ll P_X} \frac{D(P_W \| P_X)}{E[d(W, v_X) - d(W, v_W)]}.
 \end{aligned}$$

In some situations, the designer of the communication system may be interested in the sensitivity of the channel capacity to the transmission resources, or in other words, the slope of the capacity-cost function. If a closed-form expression for the capacity-cost function is not available and the capacity can only be obtained numerically, then the following result, which can be obtained by generalizing the proof of Theorem 3, is of interest. If  $P_{X_0}$  is the input distribution that achieves capacity at cost equal to  $\beta_0$ , i.e.,  $C(\beta_0) = I(X_0; Y_0)$ , then

$$C'(\beta_0) = \sup_{\substack{x \in A \\ b[x] > \beta_0}} \frac{D(P_{Y|X=x} \| P_{Y_0}) - C(\beta_0)}{b[x] - \beta_0}$$

that suggests that if  $A = R$ , and  $b[x]$  is symmetric and increasing in the positive real line, then

$$C(\beta_0) = D(P_{Y|X = +b^{-1}[\beta_0]} \| P_{Y_0}) \quad (33)$$

whenever the expression in the right side of (33) is concave in  $\beta_0$ .

#### ACKNOWLEDGMENT

The author is indebted to Bob Gallager for providing a preprint of [10], which prompted the main theme of this paper.

#### REFERENCES

- [1] R. Ash, *Information Theory*. New York: Wiley Interscience, 1965.
- [2] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
- [3] P. Bremaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer-Verlag, 1981.
- [4] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493-507, 1952.
- [5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probability*, vol. 3, pp. 146-158, Feb. 1975.
- [6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [7] M. Davis, "Capacity and cutoff rate for Poisson-type channels," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 710-715, Nov. 1980.
- [8] A. El-Gamal and T. Cover, "Multiple user information theory," *IEEE Proc.*, vol. 68, Dec. 1980, pp. 1466-1483.
- [9] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [10] R. G. Gallager, "Energy limited channels; Coding, multiaccess and spread spectrum," preprint, Nov. 1987 and in *Proc. 1988 Conf. Inform. Sci. Syst.*, p. 372, Princeton, NJ, Mar. 1988.
- [11] M. J. E. Golay, "Note on the theoretical efficiency of information reception with PPM," *Proc. IRE*, vol. 37, Sept. 1949, p. 1031.
- [12] M. Jimbo and K. Kunisawa, "An iteration method for calculating the relative capacity," *Inform. Contr.*, vol. 43, pp. 216-223, 1979.
- [13] S. Kullback, "A lower bound for discrimination information in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 126-127, Jan. 1967.
- [14] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
- [15] R. J. McEliece, *The Theory of Information and Coding: A Mathematical Framework for Communication*. Reading, MA: Addison-Wesley, 1977.
- [16] B. Meister and W. Oettli, "On the capacity of a discrete, constant channel," *Inform. Contr.*, vol. 11, pp. 341-351, 1967.
- [17] J. R. Pierce, "Optical channels: Practical limits with photon counting," *IEEE Trans. Commun.*, vol. COM-26, pp. 1819-1821, Dec. 1978.
- [18] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.
- [19] F. M. Reza, *An Introduction to Information Theory*. New York: McGraw-Hill, 1961.
- [20] H. L. Royden, *Real Analysis*. New York: MacMillan, 1968.
- [21] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, July-Oct. 1948.
- [22] A. D. Wyner, "Capacity and error exponent for the direct detection photon channel," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1449-1471, Nov. 1988.