A General Formula for Channel Capacity

Sergio Verdú, *Fellow, IEEE, and* Te Sun Han, *Fellow, IEEE*

Abstract—**A formula for the capacity of arbitrary single-user channels without feedback (not necessarily information stable, stationary, etc.) is proved. Capacity is shown to equal the supremum, over all input processes, of the input–output *inf-information rate* defined as the liminf in probability of the normalized information density. The key to this result is a new converse approach based on a simple new lower bound on the error probability of m-ary hypothesis tests among equiprobable hypotheses. A necessary and sufficient condition for the validity of the strong converse is given, as well as general expressions for ϵ-capacity.**

Index Terms—**Shannon theory, channel capacity, channel coding theorem, channels with memory, strong converse.**

I. Introduction

SHANNON'S formula [1] for channel capacity (the supremum of all rates R for which there exist sequences of codes with vanishing error probability and whose size grows with the block length n as $\exp(nR)$),

$$C = \max_X I(X;Y), \tag{1.1}$$

holds for *memoryless* channels. If the channel has memory, then (1.1) generalizes to the familiar limiting expression

$$C = \lim_{n\to\infty} \sup_{X^n} \frac{1}{n} I(X^n;Y^n). \tag{1.2}$$

However, the capacity formula (1.2) does not hold in full generality; its validity was proved by Dobrushin [2] for the class of *information stable* channels. Those channels can be roughly described as having the property that the input that maximizes mutual information and its corresponding output behave ergodically. That ergodic behavior is the key to generalize the use of the law of large numbers in the proof of the direct part of the memoryless channel coding theorem. Information stability is not a superfluous sufficient condition for the validity of (1.2).[1] Consider a binary channel where the output codeword is equal to the transmitted codeword with probability 1/2 and independent of the transmitted codeword with probability 1/2. The capacity of this channel is equal to 0 because arbitrarily small error probability is unattainable. However the right-hand side of (1.2) is equal to 1/2 bit/channel use.

The immediate question is whether there exists a completely general formula for channel capacity, which does not require any assumption such as memorylessness, information stability, stationarity, causality, etc. Such a formula is found in this paper.

Finding expressions for channel capacity in terms of the probabilistic description of the channel is the purpose of channel coding theorems. The literature on coding theorems for single-user channels is vast (cf., e.g., [4]). Since Dobrushin's information stability condition is not always easy to check for specific channels, a large number of works have been devoted to showing the validity of (1.2) for classes of channels characterized by their memory structure, such as finite-memory and asymptotically memoryless conditions. The first example of a channel for which formula (1.2) fails to hold was given in 1957 by Nedoma [5]. In order to go beyond (1.2) and obtain capacity formulas for *information unstable* channels, researchers typically considered averages of stationary ergodic channels, i.e., channels which, conditioned on the initial choice of a parameter, are information stable. A formula for averaged discrete memoryless channels was obtained by Ahlswede [6] where he realized that the Fano inequality fell short of providing a tight converse for those channels. Another class of chanels that are not necessarily information stable was studied by Winkelbauer [7]: stationary discrete regular decomposable channels with finite input memory. Using the ergodic decomposition theorem, Winkelbauer arrived at a formula for ϵ-capacity that holds for all but a countable number of values of ϵ. Nedoma [8] had shown that some stationary nonergodic channels cannot be represented as a mixture of ergodic channels; however, the use of the ergodic decomposition theorem was circumvented by Kieffer [9] who showed that Winkelbauer's capacity formula applies to all discrete stationary nonanticipatory channels. This was achieved by a converse whose proof involves Fano's and Chebyshev's inequalities plus a generalized Shannon–McMillan Theorem for periodic measures. The stationarity of the channel is a crucial assumption in that argument.

Using the Fano inequality, it can be easily shown (cf. Section III) that the capacity of every channel (defined in

Manuscript received December 15, 1992; revised June 12, 1993. This work was supported in part by the National Science Foundation under PYI Award ECSE-8857689 and by a grant from NEC. This paper was presented in part at the 1993 IEEE workshop on Information Theory, Shizuoka, Japan, June 1993.

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

T. S. Han is with the Graduate School of Information Systems, University of Electro-Communications, Tokyo 182, Japan.

IEEE Log Number 9402452.

[1] In fact, it was shown by Hu [3] that information stability is essentially equivalent to the validity of formula (1.2).

the conventional way, cf. Section II) satisfies

$$C \le \liminf_{n\to\infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n). \tag{1.3}$$

To establish equality in (1.3), the direct part of the coding theorem needs to assume information stability of the channel. Thus, the main existing results that constitute our starting point are a converse theorem (i.e., an upper bound on capacity) which holds in full generality and a direct theorem which holds for information stable channels. At first glance, this may lead one to conclude that the key to a general capacity formula is a new direct theorem which holds without assumptions. However, the foregoing example shows that the converse (1.3) is not tight in that case. Thus, what is needed is a new converse which is tight for every channel. Such a converse is the main result of this paper. It is obtained without recourse to the Fano inequality which, as we will see, cannot lead to the desired result. The proof that the new converse is tight (i.e., a general direct theorem) follows from the conventional argument once the right definition is made.

The capacity formula proved in this paper is

$$C = \sup_{\boldsymbol{X}} \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}). \tag{1.4}$$

In (1.4), $\boldsymbol{X}$ denotes an input process in the form of a sequence of finite-dimensional distributions $\boldsymbol{X} = \{X^n = (X_1^{(n)}, \cdots, X_n^{(n)})\}_{n=1}^{\infty}$. We denote by $\boldsymbol{Y} = \{Y^n = (Y_1^{(n)}, \cdots, Y_n^{(n)})\}_{n=1}^{\infty}$ the corresponding output sequence of finite-dimensional distributions induced by $\boldsymbol{X}$ via the channel $\boldsymbol{W} = \{W^n = P_{Y^n|X^n}: A^n \to B^n\}_{n=1}^{\infty}$, which is an arbitrary sequence of n-dimensional conditional output distributions from A^n to B^n, where A and B are the input and output alphabets, respectively.[2] The symbol $\underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y})$ appearing in (1.4) is the *inf-information rate* between $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is defined in [10] as the *liminf in probability*[3] of the sequence of normalized information densities $(1/n) i_{X^n W^n}(X^n; Y^n)$, where

$$i_{X^n W^n}(a^n; b^n) = \log \frac{P_{Y^n|X^n}(b^n|a^n)}{P_{Y^n}(b^n)}. \tag{1.5}$$

For ease of notation and to highlight the simplicity of the proofs, we have assumed in (1.5) and throughout the paper that the input and output alphabets are finite. However, it will be apparent from our proofs that the results of this paper do not depend on that assumption. They can be shown for channels with abstract alphabets by working with a general information density defined in the conventional way [11] as the log derivative of the conditional output measure with respect to the unconditional output measure.

The notion of inf/sup-information/entropy rates and the recognition of their key role in dealing with nonergodic/nonstationary sources are due to [10]. In particular, that paper shows that the minimum achievable source coding rate for any finite-alphabet source $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ is equal to its sup-entropy rate $\overline{\boldsymbol{H}}(\boldsymbol{X})$, defined as the limsup in probability of $(1/n) \log 1/P_{X^n}(X^n)$. In contrast to the general capacity formula presented in this paper, the general source coding result can be shown by generalizing existing proofs.

The definition of channel as a sequence of finite-dimensional conditional distributions can be found in well-known contributions to the Shannon-theoretic literature (e.g., Dobrushin [2], Wolfowitz [12, ch. 7], and Csiszár and Körner [13, p. 100]), although, as we saw, previous coding theorems imposed restrictions on the allowable class of conditional distributions. Essentially the same general channel model was analyzed in [26] arriving at a capacity formula which is not quite correct. A different approach has been followed in the ergodic-theoretic literature, which defines a channel as a conditional distribution between spaces of doubly infinite sequences.[4] In that setting (and within the domain of block coding [14]), codewords are preceded by a prehistory (a left-sided infinite sequence) and followed by a posthistory (a right-sided infinite sequence); the error probability may be defined in a worst case sense over all possible input pre- and posthistories. The channel definition adopted in this paper, namely, a sequence of finite-dimensional distributions, captures the physical situation to be modeled where block codewords are transmitted through the channel. It is possible to encompass physical models that incorporate anticipation, unlimited memory, nonstationarity, etc., because we avoid placing restrictions on the sequence of conditional distributions. Instead of taking the worst case error probability over all possible pre- and posthistories, whatever statistical knowledge is available about those sequences can be incorporated by averaging the conditional transition probabilities (and, thus, averaging the error probability) over all possible pre- and posthistories. For example, consider a simple channel with memory:

$$y_i = x_i + x_{i-1} + n_i.$$

where $\{n_i\}$ is an i.i.d. sequence with distribution P_N. The posthistory to any n-block codeword is irrelevant since this channel is causal. The conditional output distribution takes the form

$$W^n(y^n|x^n) = P_{Y_1|X_1}(y_1|x_1) \prod_{i=2}^{n} P_N(y_i - x_i - x_{i-1})$$

where the statistical information about the prehistory (summarized by the distribution of the initial state) only affects $P_{Y_1|X_1}$:

$$P_{Y_1|X_1}(y_1|x_1) = \sum_{x_0} P_N(y_1 - x_1 - x_0) P_{X_0}(x_0).$$

[2] The methods of this paper allow the study, with routine modifications, of even more abstract channels defined by arbitrary sequences of conditional output distributions, which need not map Cartesian products of the input/output alphabets. The only requirement is that the index of the sequence be the parameter that divides the amount of information in the definition of rate.

[3] If A_n is a sequence of random variables, its *liminf in probability* is the supremum of all the reals α for which $P[A_n \le \alpha] \to 0$ as $n \to \infty$. Similarly, its *limsup in probability* is the infimum of all the reals β for which $P[A_n \ge \beta] \to 0$ as $n \to \infty$.

[4] Or occasionally semi-infinite sequences, as in [9].

In this case, the choice of $P_{X_0}(x_0)$ does not affect the value of the capacity. In general, if a worst case approach is desired, an alternative to the aforementioned approach is to adopt a compound channel model [12] defined as a family of sequences of finite-dimensional distributions parametrized by the unknown initial state which belongs to an uncertainty set. That model, or the more general arbitrarily varying channel, incorporates nonprobabilistic modeling of uncertainty, and is thus outside the scope of this paper.

In Section II, we show the direct part of the capacity formula $C \geq \sup_X \underline{I}(X;Y)$. This result follows in a straightforward fashion from Feinstein's lemma [15] and the definition of inf-information rate. Section III is devoted to the proof of the converse $C \leq \sup_X \underline{I}(X;Y)$. It presents a new approach to the converse of the coding theorem based on a simple lower bound on error probability that can be seen as a natural counterpart to the upper bound provided by Feinstein's lemma. That new lower bound, along with the upper bound in Feinstein's lemma, are shown to lead to tight results on the ϵ-capacity of arbitrary channels in Section IV. Another application of the new lower bound is given in Section V: a necessary and sufficient condition for the validity of the strong converse. Section VI shows that many of the familiar properties of mutual information are satisfied by the inf-information rate, thereby facilitating the evaluation of the general formula (1.4). Examples of said evaluation for channels that are not encompassed by previous formulas can be found in Section VII.

II. Direct Coding Theorem: $C \geq \sup_X \underline{I}(X;Y)$

The conventional definition of channel capacity is (e.g., [13]) the following.

Definition 1: An (n, M, ϵ) code has block length n, M codewords, and error probability[5] not larger than ϵ. $R \geq 0$ is an *ϵ-achievable rate* if, for every $\delta > 0$, there exist, for all sufficiently large n, (n, M, ϵ) codes with rate

$$\frac{\log M}{n} > R - \delta.$$

The maximum ϵ-achievable rate is called the *ϵ-capacity* C_ϵ. The *channel capacity* C is defined as the maximal rate that is ϵ-achievable for all $0 < \epsilon < 1$. It follows immediately from the definition that $C = \lim_{\epsilon \downarrow 0} C_\epsilon$.

The basis to prove the desired lower and upper bounds on capacity are respective upper and lower bounds on the error probability of a code as a function of its size. The following classical result (Feinstein's lemma) [15] shows the existence of a code with a guaranteed error probability as a function of its size.

Theorem 1: Fix a positive integer n and $0 < \epsilon < 1$. For every $\gamma > 0$ and input distribution P_{X^n} on A^n, there exists an (n, M, ϵ) code for the transition probability $W^n = P_{Y^n|X^n}$ that satisfies

$$\epsilon \leq P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n}\log M + \gamma\right] + \exp(-\gamma n). \tag{2.1}$$

Note that Theorem 1 applies to arbitrary fixed block length and, moreover, to general random transformations from input to output, not necessarily only to transformations between nth Cartesian products of sets. However, we have chosen to state Theorem 1 in that setting for the sake of concreteness.

Armed with Theorem 1 and the definitions of capacity and inf-information rate, it is now straightforward to prove the direct part of the coding theorem.

Theorem 2: [6]

$$C \geq \sup_X \underline{I}(X;Y). \tag{2.2}$$

Proof: Fix arbitrary $0 < \epsilon < 1$ and X. We shall show that $\underline{I}(X;Y)$ is an ϵ-achievable rate by demonstrating that, for every $\delta > 0$ and all sufficiently large n, there exist $(n, M, \exp(-n\delta/4) + \epsilon/2)$ codes with rate

$$\underline{I}(X;Y) - \delta < \frac{\log M}{n} < \underline{I}(X;Y) - \frac{\delta}{2}. \tag{2.3}$$

If, in Theorem 1, we choose $\gamma = \delta/4$, then the probability in (2.1) becomes

$$P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n}\log M + \delta/4\right]$$
$$\leq P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \underline{I}(X;Y) - \delta/4\right] \leq \frac{\epsilon}{2} \tag{2.4}$$

where the second inequality holds for all sufficiently large n because of the definition of $\underline{I}(X;Y)$. In view of (2.4), Theorem 1 guarantees the existence of the desired codes. □

III. Converse Coding Theorem: $C \leq \sup_X \underline{I}(X;Y)$

This section is devoted to our main result: a tight converse that holds in full generality. To that end, we need to obtain for any arbitrary code a lower bound on its error probability as a function of its size or, equivalently, an upper bound on its size as a function of its error probability. One such bound is the standard one resulting from the Fano inequality.

Theorem 3: Every (n, M, ϵ) code satisfies

$$\log M \leq \frac{1}{1-\epsilon}[I(X^n; Y^n) + h(\epsilon)] \tag{3.1}$$

where h is the binary entropy function, X^n is the input distribution that places probability mass $1/M$ on each of the input codewords, and Y^n is its corresponding output distribution.

[5] We work throughout with average error probability. It is well known that the capacity of a single-user channel with known statistical description remains the same under the maximal error probability criterion.

[6] Whenever we omit the set over which the supremum is taken, it is understood that it is equal to the set of all sequences of finite-dimensional distributions on input strings.

Using Theorem 3, it is evident that if $R \geq 0$ is ϵ-achievable, then for every $\delta > 0$.

$$R - \delta < \frac{1}{1-\epsilon}\left[\frac{1}{n}\sup_{X^n} I(X^n;Y^n) + \frac{h(\epsilon)}{n}\right] \quad (3.2)$$

which, in turn, implies

$$R \leq \frac{1}{1-\epsilon}\operatorname*{liminf}_{n\to\infty}\frac{1}{n}\sup_{X^n} I(X^n;Y^n). \quad (3.3)$$

Thus, the general converse in (1.3) follows by letting $\epsilon \to 0$. But, as we illustrated in Section I, (1.3) is not always tight. The standard bound in Theorem 3 falls short of leading to the desired tight converse because it depends on the channel through the input–output *mutual information* (expectation of information density) achieved by the code. Instead, we need a finer bound that depends on the distribution of the information density achieved by the code, rather than on just its expectation. The following basic result provides such a bound in a form which is pleasingly dual to the Feinstein bound. As for the Feinstein bound, Theorem 4 holds not only for arbitrary fixed block length, but for an arbitrary random transformation.

Theorem 4: Every (n, M, ϵ) code satisfies

$$\epsilon \geq P\left[\frac{1}{n} i_{X^nW^n}(X^n;Y^n) \leq \frac{1}{n}\log M - \gamma\right] - \exp(-\gamma n) \quad (3.4)$$

for every $\gamma > 0$, where X^n places probability mass $1/M$ on each codeword.

Proof: Denote $\beta = \exp(-\gamma n)$. Note first that the event whose probability appears in (3.4) is equal to the set of "atypical" input–output pairs

$$L = \left\{(a^n, b^n) \in A^n \times B^n : P_{X^n|Y^n}(a^n|b^n) \leq \beta\right\}. \quad (3.5)$$

This is because the information density can be written as

$$i_{X^nW^n}(a^n;b^n) = \log\frac{P_{X^n|Y^n}(a^n|b^n)}{P_{X^n}(a^n)} \quad (3.6)$$

and $P_{X^n}(c_i) = 1/M$ for each of the M codewords $c_i \in A^n$.

We need to show that

$$P_{X^nY^n}[L] \leq \epsilon + \beta. \quad (3.7)$$

Now, denoting the decoding set corresponding to c_i by D_i and

$$B_i = \left\{b^n \in B^n : P_{X^n|Y^n}(c_i|b^n) \leq \beta\right\} \quad (3.8)$$

we can write

$$\begin{aligned} P_{X^nY^n}[L] &= \sum_{i=1}^{M} P_{X^nY^n}[(c_i, B_i)] \\ &= \sum_{i=1}^{M} P_{X^nY^n}[(c_i, B_i \cap D_i^c)] \\ &\quad + \sum_{i=1}^{M} P_{X^nY^n}[(c_i, B_i \cap D_i)] \\ &\leq \sum_{i=1}^{M}\frac{1}{M}W^n(D_i^c|c_i) + \sum_{i=1}^{M} P_{X^nY^n}[(c_i, B_i \cap D_i)] \\ &\leq \sum_{i=1}^{M}\frac{1}{M}W^n(D_i^c|c_i) + \beta P_{Y^n}\left[\bigcup_{i=1}^{M} D_i\right] \\ &\leq \epsilon + \beta \end{aligned} \quad (3.9)$$

where the second inequality is due to (3.8) and the disjointness of the decoding sets. □

Theorems 3 and 4 hold for arbitrary random transformations, in which general setting they are nothing but lower bounds on the minimum error probability of M-ary equiprobable hypothesis testing. If, in that general setting, we denote the observations by Y and the true hypothesis by X (equiprobable on $\{1,\cdots,M\}$), the M hypothesized distributions are the conditional distributions $\{P_{Y|X=i}, i = 1,\cdots,M\}$. The bound in Theorem 3 yields

$$\epsilon \geq 1 - \frac{I(X;Y) + \log 2}{\log M}.$$

A slightly weaker result is known in statistical inference as Fano's lemma [16]. The bound in Theorem 4 can easily be seen to be equivalent to the more general version

$$\epsilon \geq P\left[P_{X|Y}(X|Y) \leq \alpha\right] - \alpha$$

for arbitrary $0 \leq \alpha \leq 1$. A stronger bound which holds without the assumption of equiprobable hypothesis has been found recently in [17].

Theorem 4 gives a family (parametrized by γ) of lower bounds on the error probability. To obtain the best bound, we simply maximize the right-hand side of (3.4) over γ. However, a judicious, if not optimum, choice of γ is sufficient for the purposes of proving the general converse.

Theorem 5:

$$C \leq \sup_X \underline{I}(X:Y). \quad (3.10)$$

Proof: The intuition behind the use of Theorem 4 to prove the converse is very simple. As a shorthand, let us refer to a sequence of codes with vanishingly small error probability (i.e., a sequence of (n, M, ϵ_n) codes such that $\epsilon_n \to 0$) as a *reliable code sequence*. Also, we will say that the *information spectrum* of a code (a term coined in [10]) is the distribution of the normalized information density evaluated with the input distribution X^n that places equal probability mass on each of the codewords of the code. Theorem 4 implies that if a reliable code sequence has rate R, then the mass of its information spectrum lying strictly to the left of R must be asymptotically negligible.

In other words, $R \le \underline{I}(X;Y)$ where X corresponds to the sequence of input distributions generated by the sequence of codebooks.

To formalize this reasoning, let us argue by contradiction and assume that for some $\rho > 0$,

$$C = \sup_X \underline{I}(X;Y) + 3\rho. \tag{3.11}$$

By definition of capacity, there exists a reliable code sequence with rate

$$\frac{\log M}{n} > C - \rho. \tag{3.12}$$

Now, letting X^n be the distribution that places probability mass $1/M$ on the codewords of that code, Theorem 4 (choosing $\gamma = \rho$), (3.11) and (3.12) imply that the error probability must be lower bounded by

$$\epsilon_n \ge P\left[\frac{1}{n} i_{X^nW^n}(X^n;Y^n) \le \sup_X \underline{I}(X;Y) + \rho\right] - \exp(-n\rho). \tag{3.13}$$

But, by definition of $\underline{I}(X;Y)$, the probability on the right-hand side of (3.13) cannot vanish asymptotically, thereby contradicting the fact that $\epsilon_n \to 0$. □

Besides the behavior of the information spectrum of a reliable code sequence revealed in the proof of Theorem 5, it is worth pointing out that the information spectrum of any code places no probability mass above its rate. To see this, simply note that (3.6) implies

$$P\left[\frac{1}{n} i_{X^nW^n}(X^n;Y^n) \le \frac{1}{n}\log M\right] = 1. \tag{3.14}$$

Thus, we can conclude that the normalized information density of a reliable code sequence converges in probability to its rate. For finite-input channels, this implies [10, Lemma 1] the same behavior for the sequence of normalized mutual informations, thereby yielding the classical bound (1.3). However, that bound is not tight for information unstable channels because, in that case, the mutual information is maximized by input distributions whose information spectrum does not converge to a single point mass (unlike the behavior of the information spectrum of a reliable code sequence).

Upon reflecting on the proofs of the general direct and converse theorems presented in Sections II and III, we can see that those results follow from asymptotically tight upper and lower bounds on error probability, and are decoupled from ergodic results such as the law of large numbers or the asymptotic equipartition property. Those ergodic results enter in the picture only as a way to particularize the general capacity formula to special classes of channels (such as memoryless or information stable channels) so that capacity can be written in terms of the mutual information rate.

Unlike the conventional approach to the converse coding theorem (Theorem 3), Theorem 4 can be used to provide a formula for ϵ-capacity as we show in Section IV. Another problem where Theorem 4 proves to be the key result is that of combined source/channel coding [18]. It turns out that when dealing with arbitrary sources and channels, the separation theorem may not hold because, in general, it could happen that a source is transmissible over a channel even if the minimum achievable source coding rate (sup-entropy rate) exceeds the channel capacity. Necessary and sufficient conditions for the transmissibility of a source over a channel are obtained in [18].

Definition 1 is the conventional definition of channel capacity (cf. [15] and [13]) where codes are required to be reliable for *all* sufficiently large block length. An alternative, more optimistic, definition of capacity can be considered where codes are required to be reliable only infinitely often. This definition is less appealing in many practical situations because of the additional uncertainty in the favorable block lengths. Both definitions turn out to lead to the same capacity formula for specific channel classes such as discrete memoryless channels [13]. However, in general, both quantities need not be equal, and the optimistic definition does not appear to admit a simple general formula such as the one in (1.4) for the conventional definition. In particular, the optimistic capacity need not be equal to the supremum of sup-information rates. See [18] for further characterization of this quantity.

The conventional definition of capacity may be faulted for being too conservative in those rare situations where the maximum amount of reliably transmissible information does not grow linearly with block length, but, rather, as $O(b(n))$. For example, consider the case $b(n) = n + n\sin(\alpha n)$. This can be easily taken into account by "seasonal adjusting:" substitution of n by $b(n)$ in the definition of rate and in all previous results.

IV. ϵ-Capacity

The fundamental tools (Theorems 1 and 4) we used in Section III to prove the general capacity formula are used in this section to find upper and lower bounds on C_ϵ, the ϵ-capacity of the channel, for $0 < \epsilon < 1$. These bounds coincide at the points where the ϵ-capacity is a continuous function of ϵ.

Theorem 6: For $0 < \epsilon < 1$, the ϵ-capacity C_ϵ satisfies

$$C_\epsilon \le \sup_X \sup\{R: F_X(R) \le \epsilon\} \tag{4.1}$$

$$C_\epsilon \ge \sup_X \sup\{R: F_X(R) < \epsilon\} \tag{4.2}$$

where $F_X(R)$ denotes the limit of cumulative distribution functions

$$F_X(R) = \limsup_{n\to\infty} P\left[\frac{1}{n} i_{X^nW^n}(X^n,Y^n) \le R\right]. \tag{4.3}$$

The bounds (4.1) and (4.2) hold with equality, except possibly at the points of discontinuity of C_ϵ, of which there are, at most, countably many.

Proof: To show (4.1), select an ϵ-achievable rate R and fix an arbitrary $\delta > 0$. We can find a sequence of (n, M, ϵ) codes such that for all sufficiently large n,

$$\frac{1}{n}\log M > R - \delta. \tag{4.4}$$

If we apply Theorem 4 to those codes, and we let X^n distribute its probability mass evenly on the nth codebook, we obtain

$$\epsilon \geq P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq \frac{1}{n}\log M - \delta\right] - \exp(-\delta n)$$

$$\geq P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq R - 2\delta\right] - \exp(-\delta n). \tag{4.5}$$

Since (4.5) holds for all sufficiently large n and every $\delta > 0$, we must have

$$F_X(R - 2\delta) \leq \epsilon, \qquad \text{for all } \delta > 0 \tag{4.6}$$

but R satisfies (4.6) if and only if $R \leq \sup\{R\colon F_X(R) \leq \epsilon\}$. Concluding, any ϵ-achievable rate is upper bounded by the right-hand side of (4.1), as we wanted to show.

In order to prove the direct part (4.2), we will show that for every X, any R belonging to the set in the following right-hand side is ϵ-achievable:

$$\{R\colon F_X(R - \delta) < \epsilon, \qquad \text{for all } \delta > 0\}$$

$$\subset \left\{R\colon \quad \text{for all } \delta > 0,\ P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq R - \delta\right]\right.$$

$$\left.\leq \epsilon - \exp(-\delta n) \text{ if } n > n_0\right\}. \tag{4.7}$$

Theorem 1 ensures the existence (for any $\delta > 0$) of a sequence of codes with rate

$$R - 3\delta \leq \frac{1}{n}\log M \leq R - 2\delta$$

and error probability not exceeding

$$\exp(-\delta n) + P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq \frac{1}{n}\log M + \delta\right]$$

$$\leq \exp(-\delta n) + P\left[\frac{1}{n} i_{X^nW^n}(X^n, Y^n) \leq R - \delta\right]$$

$$\leq \epsilon$$

for all sufficiently large n. Thus, R is ϵ-achievable.

It is easy to see that the bounds are tight except at the points of discontinuity of C_ϵ. Let $u(\epsilon)$ and $l(\epsilon)$ denote the right-hand sides of (4.1) and (4.2), respectively. Since $u(\epsilon)$ is monotone nondecreasing, the set $D \subset (0, 1)$ of ϵ at which it is discontinuous is, at most, countable. Select any $\epsilon \in (0, 1) - D$ and a strictly increasing sequence $(\epsilon_1, \epsilon_2, \cdots)$ in $(0, 1)$ converging to ϵ. Since $F_X(R)$ is nondecreasing, we have

$$\sup\{R\colon F_X(R) < \epsilon\} = \sup_i \sup\{R\colon F_X(R) \leq \epsilon_i\}$$

from which it follows that

$$l(\epsilon) = \sup_X \sup_i \sup\{R\colon F_X(R) \leq \epsilon_i\}$$

$$= \sup_i \sup_X \sup\{R\colon F_X(R) \leq \epsilon_i\}$$

$$= \sup_i u(\epsilon_i) = u(\epsilon)$$

where the last equality holds because $u(\cdot)$ is continuous nondecreasing at ϵ. □

In the special case of stationary discrete channels, the functional in (4.1) boils down to the quantile introduced in [7] to determine ϵ-capacity, except for a countable number of values of ϵ. The ϵ-capacity formula in [7] was actually proved for a class of discrete stationary channels (so-called regular decomposable channels) that includes ergodic channels and a narrow class of nonergodic channels. The formula for the capacity of discrete stationary nonanticipatory channels given in [9] is the limit as $\epsilon \to 0$ of the right-hand side of (4.2) specialized to that particular case.

The inability to obtain an expression for ϵ-capacity at its points of discontinuity is a consequence of the definition itself rather than of our methods of analysis. In fact, it is easily checked by slightly modifying the proof of Theorem 6 that (4.1) would hold with equality for *all* $0 \leq \epsilon < 1$ had ϵ-achievable rates been defined in a slightly different (and more regular) way, by requiring sequences of codes with both rate and error probability arbitrarily close to R and ϵ, respectively. More precisely, consider an alternative definition of R as an ϵ-achievable rate ($0 \leq \epsilon < 1$) when there exists a sequence of (n, M, ϵ_n) codes with

$$\liminf_{n\to\infty} \frac{\log M}{n} \geq R$$

and

$$\limsup_{n\to\infty} \epsilon_n \leq \epsilon.$$

With this definition, the resulting C_ϵ would be the right-continuous version of the conventional ϵ-capacity, (4.1) would hold with equality for all $0 \leq \epsilon < 1$, and the channel capacity could be written as

$$C_0 = \lim_{\epsilon \downarrow 0} C_\epsilon = \sup_X \underline{I}(X; Y) = \sup_X \sup\{R\colon F_X(R) = 0\}.$$

A separate definition would then be needed for zero-error capacity—not a bad idea since it is a completely different problem.

V. Strong Converse Condition

Definition 2: A channel with capacity C is said to satisfy the *strong converse* if for every $\delta > 0$ and every sequence of (n, M, λ_n) codes with rate

$$\frac{\log M}{n} > C + \delta$$

it holds that $\lambda_n \to 1$ as $n \to \infty$.

This concept was championed by Wolfowitz [19], [20], and it has received considerable attention in information theory. In this section, we prove that it is intimately related to the form taken by the capacity formula established in this paper.

Consider the *sup-information rate* $\bar{I}(X;Y)$, whose definition is dual to that of the inf-information rate $\underline{I}(X;Y)$, that is, $\bar{I}(X;Y)$ is the limsup in probability (cf. footnote 3) of the normalized information density due to the input X. Then, Theorem 4 plays a key role in the proof of the following result.

Theorem 7: For any channel, the following two conditions are equivalent:

1) The channel satisfies the strong converse.

2) $\sup_X \underline{I}(X;Y) = \sup_X \bar{I}(X;Y)$.

Proof: It is shown in the proof of [10, Theorem 8] that the capacity is lower bounded by $C \geq \sup_X \bar{I}(X;Y)$ if the channel satisfies the strong converse. Together with the capacity formula (1.4) and the obvious inequality $\underline{I}(X;Y) \leq \bar{I}(X;Y)$, we conclude that condition 1) implies condition 2).

To show the reverse implication, fix $\delta > 0$, and select any sequence of (n, M, λ_n) codes that satisfy

$$\frac{\log M}{n} > C + \delta$$

for all sufficiently large n. Once we apply Theorem 4 to this sequence of codes, we get (with $\gamma = \delta/2$)

$$\begin{aligned}\lambda_n &\geq P\left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \leq \frac{1}{n}\log M - \delta/2\right] \\ &\quad - \exp(-\delta n/2) \\ &\geq P\left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \leq C + \delta/2\right] - \exp(-\delta n/2)\end{aligned} \tag{5.1}$$

for all sufficiently large n. But since condition 2) implies $C = \sup_X \bar{I}(X;Y)$, the probability on the right-hand side of (5.1) must go to 1 as $n \to \infty$ by definition of $\bar{I}(X;Y)$. Thus $\lambda_n \to 1$, as we wanted to show. □

Due to its full generality, Theorem 7 provides a powerful tool for studying the strong converse property.

The problem of approximation theory of channel output statistics introduced in [10] led to the introduction of the concept of channel *resolvability*. It was shown in [10] that the resolvability S of any finite-input channel is given by the dual to (1.4)

$$S = \sup_X \bar{I}(X;Y). \tag{5.2}$$

It was shown in [10] that if a finite-input channel satisfies the strong converse, then its resolvability and capacity coincide (and the conventional capacity formula (1.2) holds). We will next show as an immediate corollary to Theorem 7 that for any finite-input channel, the validity of the strong converse is not only sufficient, but also necessary for the equality $S = C$ to hold.

Corollary: If the input alphabet is finite, then the following two conditions are equivalent.

1) The channel satisfies the strong converse.

2) $C = S = \lim_{n\to\infty} \sup_{X^n} (1/n) I(X^n; Y^n)$.

Proof: Because of (1.4) and (5.2), all we need is to show the second equality in condition 2) when $\sup_X \underline{I}(X;Y) = \sup_X \bar{I}(X;Y)$. This has been shown in the proof of [10, Theorem 7]. □

Wolfowitz [20] defined capacity only for channels that satisfy the strong converse, and referred to the conventional capacity of Definition 1 (which is always defined) as *weak capacity*. The corollary shows that the *strong capacity* of finite-input channels is given by formula (1.2). It should be cautioned that the validity of the capacity formula in (1.2) is not sufficient for the strong converse to hold. In view of Theorem 7, this means that there exist channels for which

$$\begin{aligned}C &= \sup_X \underline{I}(X;Y) \\ &= \lim_{n\to\infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n) \\ &< \sup_X \bar{I}(X;Y).\end{aligned}$$

For example, consider a channel with alphabets $A = B = \{0, 1, \alpha, \beta\}$ and transition probability

$$W^n(x_1, \cdots, x_n | x_1, \cdots, x_n) = \begin{cases} 1 & \text{if } (x_1, \cdots, x_n) \in D_n \\ 0.01 & \text{if } (x_1, \cdots, x_n) \notin D_n \end{cases}$$

$$W^n(\alpha, \cdots, \alpha | x_1, \cdots, x_n) = 0.99 \quad \text{if } (x_1, \cdots, x_n) \notin D_n.$$

where $D_n = \{0,1\}^n \cup (\alpha, \ldots \alpha)$. Then

$$C = \underline{I}(X_1; Y_1) = \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n) = 1 \text{ bit}$$

and

$$\sup_X \bar{I}(X;Y) = \bar{I}(X_2; Y_2) = 2 \text{ bits}$$

where X_1 is i.i.d. equally likely on $\{0,1\}$ and X_2 is i.i.d. equally likely on $\{0, 1, \alpha, \beta\}$.

VI. Properties of Inf-Information Rate

Many of the familiar properties satisfied by mutual information turn out to be inherited by the inf-information rate. Those properties are particularly useful in the computation of $\sup_X \underline{I}(X;Y)$ for specific channels. Here, we will show a sample of the most important properties. As is well known, the nonnegativity of divergence (itself a consequence of Jensen's inequality) is the key to the proof of many of the mutual information properties. In the present context, the property that plays a major role is unrelated to convex inequalities. That property is the

nonnegativity of the *inf-divergence rate* $\underline{D}(U\|V)$ defined for two arbitrary processes U and V as the liminf in probability of the sequence of log-likelihood ratios

$$\frac{1}{n}\log\frac{P_{U^n}(U^n)}{P_{V^n}(U^n)}.$$

The *sup-entropy rate* $\overline{H}(Y)$ and *inf-entropy rate* $\underline{H}(Y)$ introduced in [10] are defined as the limsup and liminf, respectively, in probability of the normalized entropy density

$$\frac{1}{n}\log\frac{1}{P_{Y^n}(Y^n)}.$$

Analogously, the *conditional sup-entropy rate* $\overline{H}(Y|X)$ is the limsup in probability (according to $\{P_{X^nY^n}\}$) of

$$\frac{1}{n}\log\frac{1}{P_{Y^n|X^n}(Y^n|X^n)}.$$

Theorem 8: An arbitrary sequence of joint distributions (X, Y) satisfies

a) $\underline{D}(X\|Y) \geq 0$

b) $\underline{I}(X;Y) = \underline{I}(Y;X)$

c) $\underline{I}(X;Y) \geq 0$

d) $\underline{I}(X;Y) \leq \underline{H}(Y) - \underline{H}(Y|X)$

$\underline{I}(X;Y) \leq \overline{H}(Y) - \overline{H}(Y|X)$

$\underline{I}(X;Y) \geq \underline{H}(Y) - \overline{H}(Y|X)$

e) $0 \leq \overline{H}(Y) < \log|B|$

f) $\underline{I}(X,Y;Z) \geq \underline{I}(X;Z)$

g) If $\overline{I}(X;Y) = \underline{I}(X;Y)$ and the input alphabet is finite, then $\underline{I}(X;Y) = \lim_{n\to\infty}(1/n)I(X^n;Y^n)$

h) $\underline{I}(X;Y) \leq \liminf_{n\to\infty}(1/n)I(X^n;Y^n)$.

Proof: Property a) holds because, for every $\delta > 0$.

$$P\left[\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{Y^n}(X^n)} \leq -\delta\right] = \sum_{x^n:\, P_{X^n}(x^n)\leq P_{Y^n}(x^n)\exp(-\delta n)} P_{X^n}(x^n) \leq \exp(-\delta n). \tag{6.1}$$

Property b) is an immediate consequence of the definition. Property c) holds because the inf-information rate is equal to the inf-divergence rate between the processes (X,Y) and $(\overline{X},\overline{Y})$, where $\overline{X}$ and $\overline{Y}$ are independent and have the same individual statistics as X and Y, respectively.

The inequalities in d) follow from

$$i_{X^nW^n}(X^n,Y^n) = \log\frac{1}{P_{Y^n}(Y^n)} - \log\frac{1}{W^n(Y^n|X^n)} \tag{6.2}$$

and the fact that the liminf in probability of a sequence of random variables $U_n + V_n$ is upper (resp. lower) bounded by the liminf in probability of U_n plus the limsup (resp. liminf) in probability of V_n.

Property e) follows from the fact that $\overline{H}(Y)$ is the minimum achievable fixed-length source coding rate for Y [10].

To show f), note first that Kolmogorov's identity holds for information densities (not just their expectations):

$$\frac{1}{n}\log\frac{P_{Z^n|X^nY^n}(Z^n|X^n,Y^n)}{P_{Z^n}(Z^n)} = \frac{1}{n}\log\frac{P_{Z^n|X^n}(Z^n|X^n)}{P_{Z^n}(Z^n)} + \frac{1}{n}\log\frac{P_{Z^n|X^nY^n}(Z^n|X^n,Y^n)}{P_{Z^n|X^n}(Z^n|X^n)}. \tag{6.3}$$

Property f) then follows because of the nonnegativity of the liminf in probability of the second normalized information density on the right-hand side of (6.3).

Property g) is [10, Lemma 1].

To show h), let us assume that $\underline{I}(X;Y)$ is finite; otherwise, the result follows immediately. Choose an arbitrarily small $\gamma > 0$ and write the mutual information as

$$\begin{aligned}\frac{1}{n}I(X^n;Y^n) &= E\left[\frac{1}{n}i_{X^nW^n}(X^n,Y^n)\right]\\ &= E\left[\frac{1}{n}i_{X^nW^n}(X^n,Y^n)1\{i_{X^nW^n}(X^n,Y^n)\leq 0\}\right]\\ &\quad + E\left[\frac{1}{n}i_{X^nW^n}(X^n,Y^n)\cdot 1\left\{0 < \frac{1}{n}i_{X^nW^n}(X^n,Y^n) \leq \underline{I}(X{:}Y) - \gamma\right\}\right]\\ &\quad + E\left[\frac{1}{n}i_{X^nW^n}(X^n,Y^n)\cdot 1\left\{\underline{I}(X;Y) - \gamma \leq \frac{1}{n}i_{X^nW^n}(X^n,Y^n)\right\}\right].\end{aligned}$$

The first term is lower bounded by $-(\log e)/(en)$ (e.g., [21]); therefore, it vanishes as $n\to\infty$. By definition of $\underline{I}(X;Y)$, the second term vanishes as well, and the third term is lower bounded by $\underline{I}(X;Y) - \gamma$. □

Theorem 9 (Data Processing Theorem): Suppose that for every n, X_1^n and X_3^n are conditionally independent given X_2^n. Then

$$\underline{I}(X_1;X_3) \leq \underline{I}(X_1;X_2). \tag{6.4}$$

Proof: By Theorem 8f), we get

$$\underline{I}(X_1;X_3) \leq \underline{I}(X_1;X_2,X_3) = \underline{I}(X_1;X_2) \tag{6.5}$$

where the equality holds because $\underline{I}(X_1;X_2,X_3)$ is the liminf in probability of [cf. (6.3)]

$$\begin{aligned}\frac{1}{n}\log\frac{P_{X_1^n|X_2^nX_3^n}(X_1^n|X_2^n,X_3^n)}{P_{X_1^n}(X_1^n)} &= \frac{1}{n}\log\frac{P_{X_1^n|X_2^n}(X_1^n|X_2^n)}{P_{X_1^n}(X_1^n)} + \frac{1}{n}\log\frac{P_{X_1^n|X_2^nX_3^n}(X_1^n|X_2^n,X_3^n)}{P_{X_1^n|X_2^n}(X_1^n|X_2^n)}\\ &= \frac{1}{n}\log\frac{P_{X_1^n|X_2^n}(X_1^n|X_2^n)}{P_{X_1^n}(X_1^n)}.\end{aligned} \tag{6.6}$$

□

Theorem 10 (Optimality of Independent Inputs): If a discrete channel is memoryless, i.e., $P_{Y^n|X^n} = W^n = \prod_{i=1}^n W_i$, for all n, then, for any input X and the corresponding output[7] Y,

$$\underline{I}(X;Y) \le \underline{I}(\bar{X};\bar{Y}) \tag{6.7}$$

where $\bar{Y}$ is the output due to $\bar{X}$, which is an independent process with the same first-order statistics as X, i.e., $P_{\bar{X}^n} = \prod_{i=1}^n P_{X_i}$.

Proof: Let α denote the liminf in probability (according to $\{P_{X^nY^n}\}$) of the sequence

$$Z_n = \frac{1}{n} i_{\bar{X}^n W^n}(X^n, Y^n) = \frac{1}{n}\sum_{i=1}^n \log \frac{W_i(Y_i|X_i)}{P_{Y_i}(Y_i)}. \tag{6.8}$$

We will show that

$$\alpha \ge \underline{I}(X;Y) \tag{6.9}$$

and

$$\alpha \le \underline{I}(\bar{X};\bar{Y}). \tag{6.10}$$

To show (6.9), we can write (6.8) as

$$\frac{1}{n}\sum_{i=1}^n \log \frac{W_i(Y_i|X_i)}{P_{Y_i}(Y_i)} = \frac{1}{n}\log \frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)} + \frac{1}{n}\log \frac{P_{Y^n}(Y^n)}{\prod_{i=1}^n P_{Y_i}(Y_i)} \tag{6.11}$$

where the liminf in probability of the first term on the right-hand side is $\underline{I}(X;Y)$ and the liminf in probability of the second term is nonnegative owing to Theorem 8a).

To show (6.10), note first (from the independence of the sequence $(\bar{X}_i, \bar{Y}_i)$, the Chebyshev inequality, and the discreteness of the alphabets) that

$$\begin{aligned}\underline{I}(\bar{X};\bar{Y}) &= \liminf_{n\to\infty} E[Z_n] \\ &= \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n I(X_i;Y_i).\end{aligned} \tag{6.12}$$

Then (6.10) will follow once we show that α, the liminf in probability of Z_n, cannot exceed (6.12). To see this, let us argue by contradiction and assume otherwise, i.e., for some $\gamma > 0$,

$$P\left[Z_n > \underline{I}(\bar{X};\bar{Y}) + \gamma\right] \to 1. \tag{6.13}$$

But for every n, we have

$$\begin{aligned}E[Z_n] &\ge E[Z_n 1\{Z_n \le 0\}] + \left(\liminf_{n\to\infty} E[Z_n] + \gamma\right) \\ &\quad \cdot P\left[Z_n > \underline{I}(\bar{X};\bar{Y}) + \gamma\right] \\ &\le -\frac{1}{n} e^{-1}\log e + \left(\liminf_{n\to\infty} E[Z_n] + \gamma\right) \\ &\quad \cdot P\left[Z_n > \underline{I}(\bar{X};\bar{Y}) + \gamma\right].\end{aligned} \tag{6.14}$$

[7]As throughout the paper, we allow that individual distributions depend on the block length, i.e., $X^n = (X_1^{(n)}, \cdots, X_n^{(n)})$, $W^n = \prod_{i=1}^n W_i^{(n)}$, etc. However, we drop the explicit dependence on (n) to simplify notation.

The second inequality is a result of

$$\begin{aligned}&nE[Z_n 1\{Z_n \le 0\}] \\ &\quad = E[g(X^n,\bar{Y}^n)\log g(X^n,\bar{Y}^n) 1\{g(X^n,\bar{Y}^n) \le 1\}] \\ &\quad \ge e^{-1}\log e^{-1}\end{aligned}$$

where $\bar{Y}^n$ is independent of X^n and

$$g(x^n, y^n) = \prod_{i=1}^n W_i(y_i|x_i)/P_{Y_i}(y_i).$$

Finally, note that (6.13) and (6.14) are incompatible for sufficiently large n. □

Analogous proofs can be used to show corresponding properties for the sup-information rate $\bar{I}(X;Y)$ arising in the problem of channel resolvability [10].

VII. Examples

As an example of a simple channel which is not encompassed by previously published formulas, consider the following binary channel.

Let the alphabets be binary $A = B = \{0,1\}$, and let every output be given by

$$Y_i = X_i + Z_i \tag{7.1}$$

where addition is modulo-2 and Z is an *arbitrary* binary random process independent of the input X. The evaluation of the general capacity formula yields

$$\sup_X \underline{I}(X;Y) = \log 2 - \bar{H}(Z) \tag{7.2}$$

where $\bar{H}(Z)$ is the sup-entropy rate of the additive noise process Z (cf. definition in Section VI). A special case of formula (7.2) was proved by Parthasarathy [22] in the context of stationary noise processes in a form which, in lieu of the sup-entropy rate, involves the supremum over the entropies of almost every ergodic component of the stationary noise.

In order to verify (7.2), we note first that, according to properties d) and e) in Theorem 8, every X satisfies

$$\underline{I}(X;Y) \le \log 2 - \bar{H}(Y|X). \tag{7.3}$$

Moreover, because of the symmetry of the channel, $\bar{H}(Y|X)$ does not depend on X. To see this, note that the distribution of $\log P_{Y^n|X^n}(Y^n[a^n]\|a^n)$ is independent of a^n when $Y^n[a^n]$ is distributed according to the conditional distribution $P_{Y^n|X^n=a^n}$. Thus, we can compute $\bar{H}(Y|X)$ with an arbitrary X. In particular, we can let the input be equal to the all-zero sequence, yielding $\bar{H}(Y|X) = \bar{H}(Z)$. To conclude the verification of (7.2), it is enough to notice that (7.3) holds with equality when X is equally likely Bernoulli.

Let us examine several examples of the computation of (7.2).

1) If the process Z is Bernoulli with parameter p, then the channel is a stationary memoryless binary symmetric channel. By the weak law of large numbers,

$(1/n)\log P_{Z^n}(Z^n)$ converges in probability to its mean. Thus, $\bar{H}(\mathbf{Z}) = h(p) = p\log(1/p) + (1-p)\log(1/1-p)$. More generally, if $\mathbf{Z}$ is stationary ergodic, then the Shannon–MacMillan theorem can be used to conclude that $\bar{H}(\mathbf{Z})$ is the entropy rate of the process.

2) Let $\mathbf{Z}$ be an all-zero sequence with probability β and Bernoulli (with parameter p) with probability $1-\beta$. In other words, the channel is either noiseless or a BSC with crossover probability p for the whole duration of the transmission (cf. [10]). The sequence of random variables $(1/n)\log(1/P_{Z^n}(Z^n))$ converges to atoms 0 and $h(p)$ with respective masses β and $1-\beta$. Thus, the minimum achievable source coding rate for $\mathbf{Z}$ is $\bar{H}(\mathbf{Z}) = h(p)$, and the channel capacity is $1-h(p)$ bits. This illustrates that the definitions of minimum source coding rate and channel capacity are based on the worst case in the sense that they designate the best rate for which arbitrarily high reliability is guaranteed regardless of the ergodic mode in effect. Universal coding [23] shows that it is possible to encode $\mathbf{Z}$ at a rate that will be either 0 or $h(p)$ depending on which mode is in effect. Thus, even though no code with a fixed rate lower than $h(p)$ will have arbitrarily small error probability, there are reliable codes with *expected* rate equal to $(1-\beta)h(p)$. Can we draw a parallel conclusion with channel capacity? At first glance, it may seem that the answer is negative because the channel encoder (unlike the source encoder or the channel decoder) cannot learn the ergodic mode in effect. However, it is indeed possible to take into account the probabilities of each ergodic mode and maximize the *expected* rate of information transfer. This is one of the applications of broadcast channels suggested by Cover [24]. The encoder chooses a code that enables reliable transmission at rates R_1 with the perfect channel and R_2 with the BSC, where (R_1, R_2) belongs to the capacity region of the corresponding broadcast channel. In addition, a preamble can be added (without penalty on the transmission rate) so that the decoder learns which channel is in effect. As the capacity result shows, we can choose $R_1 = R_2 = 1 - h(p)$. However, in some situations, it may make more sense to maximize the expected rate $\beta R_1 + (1-\beta)R_2$ instead of the worst case $\min\{R_1, R_2\}$. The penalty incurred because the encoder is not informed of the ergodic mode in effect is that the maximum expected rate is strictly smaller than the average of the individual capacities and is equal to [24]

$$\max_{0\le\alpha\le 1}\{1 - h(\alpha + p - 2\alpha p) + \beta h(\alpha)\}.$$

In general, the problem of maximizing the expected rate (for an arbitrary mixture distribution) of a channel with K ergodic modes is equivalent to finding the capacity region of the corresponding K-user broadcast channel (still an open problem, in general).

3) if $\mathbf{Z}$ is a homogeneous Markov chain (not necessarily stationary or ergodic), then $\bar{H}(\mathbf{Z})$ is equal to zero if the chain is nonergodic, and to the conventional conditional entropy of the steady-state chain if the chain is ergodic. This result is easy to generalize to nonbinary chains, where the sup-entropy rate is given by the largest conditional entropy (over all steady-state distributions).

4) If $\mathbf{Z}$ is an independent nonstationary process with $P[Z_i = 1] = \delta_i$, then

$$\bar{H}(\mathbf{Z}) = \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} h(\delta_i).$$

To see this, consider first the case where the sequence δ_i takes values on a finite set. Then the result follows upon the application of the weak law of large numbers to each of the subsequences with a common crossover probability. In general, the result can be shown by partitioning the unit interval into arbitrarily short segments and considering the subsequences corresponding to crossover probabilities within each segment.

5) Define the following nonergodic nonstationary $\mathbf{Z}$: time is partitioned in blocks of length 1, 1, 2, 4, 8,$\cdots$, and we label those blocks starting with $k = 0$. Thus, the length of the kth block is 1 for $k = 0$ and 2^{k-1} for $k = 1, 2, \cdots$. Note that the cumulative length up to and including the kth block is 2^k. The process is independent from block to block. In each block, the process is equal to the all-zero vector with probability $1/2$ and independent equally likely with probability $1/2$. In other words, the channel is either a BSC with crossover probability $1/2$ or a noiseless BSC according to a switch which is equally likely to be in either position and may change position only after times 1, 2, 4, $8, \cdots$. We will sketch a proof of $\bar{H}(\mathbf{Z}) = 1$ bit (and, thus, the capacity is zero) by considering the sequence of normalized log-likelihoods for block lengths $n = 2^k$:

$$W_k = 2^{-k}\log 1/P_{Z^{2^k}}(Z^{2^k}).$$

This sequence of random variables satisfies the dynamics

$$W_{k+1} = (W_k + L_k + \Delta_k)/2$$

where the random variables $\{L_k + \Delta_k\}$ are independent, L_k is equal to 0 with probability $1/2$ and equal to 1 bit with probability $1/2$, and Δ_k is a positive random variable which is upper bounded by 2^{1-k} bit (and is dependent on L_k). The asymptotic behavior of W_k is identical to the case where $\Delta_k = 0$ because the random variable $\sum_{i=1}^{k} 2^{-i}\Delta_{k-i}$ converges to zero almost surely as $k \to \infty$. Then, it can be checked [25] that W_k converges in law to a uniform distribution on $[0, 1]$, and thus, its limsup in probability is equal to 1 bit. Applying the formula for ϵ-capacity in Theorem 6, we obtain $C_\epsilon = \epsilon$.

VIII. Conclusion

A new approach to the converse of the channel coding theorem, which can be considered to be the dual of Feinstein's lemma, has led to a completely general formula for channel capacity. The simplicity of this approach should not be overshadowed by its generality.

No results on convergence of time averages (such as ergodic theory) enter the picture in order to derive the general capacity formula. It is only in the particularization of that formula to specific channels that we need to use the law of large numbers (or, in general, ergodic theory).

The utility of inf-information rate goes beyond the fact that it is the "right" generalization of the conventional mutual information rate. There are cases where, even if conventional expressions such as (1.2) hold, it is advantageous to work with inf-information rates. For example, in order to show the achievability result $C \geq \alpha$, it is enough to show that $\underline{I}(\boldsymbol{X};\boldsymbol{Y}) \geq \alpha$ for some input process, to which end it is not necessary to show convergence of the information density to its expected value.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July–Oct. 1948.

[2] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Amer. Math. Soc. Trans.*, vol. 33, pp. 323–438, AMS, Providence, RI, 1963.

[3] G. D. Hu, "On Shannon theorem and its converse for sequences of communication schemes in the case of abstract random variables," in *Proc. 3rd Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes*. Prague: Czechoslovak Acad. Sci., 1964, pp. 285–333.

[4] R. M. Gray and L. D. Davisson, *Ergodic and Information Theory*. Stroudsbourg, PA: Dowden, Hutchinson & Ross, 1977.

[5] J. Nedoma, "The capacity of a discrete channel," in *Proc. 1st Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes*, Prague, 1957, pp. 143–181.

[6] R. Ahlswede, "The weak capacity of averaged channels," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 11, pp. 61–73, 1968.

[7] K. Winkelbauer, "On the coding theorem for decomposable discrete information channels I," *Kybernetika*, vol. 7, no. 2, pp. 109–123, 1971.

[8] J. Nedoma, "On nonergodic channels," in *Proc. 2nd Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes*, Prague, 1960, pp. 363–395.

[9] J. C. Kieffer, "A general formula for the capacity of stationary nonanticipatory channels," *Inform. Contr.*, vol. 26, pp. 381–391, 1974.

[10] T. S. Han and S. Verdu, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752 772, May 1993.

[11] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[12] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. New York: Springer, 1978.

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

[14] R. M. Gray and D. S. Ornstein, "Block coding for discrete stationary d-bar continuous noisy channels," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 292–306, May 1979.

[15] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inform. Theory*, vol. IT-4, pp. 2–22, 1954.

[16] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer, 1986.

[17] H. V. Poor and S. Verdú, "A lower bound on the probability of error in multihypothesis testing," in *Proc. 1993 Allerton Conf. Commun., Contr., Computing*, Monticello, IL, pp. 758–759, Sept. 1993.

[18] S. Vembu, S. Verdú, and Y. Steinberg, "The joint source-channel separation theorem revisited," IEEE Trans. Inform. Theory, vol. 41, to appear, 1995.

[19] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, pp. 591–606, Dec. 1957.

[20] ——, "On channels without capacity," *Inform. Contr.*, vol. 6, pp. 49–54, 1963.

[21] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.

[22] K. R. Parthasarathy, "Effective entropy rate and transmission of information through channels with additive random noise," *Sankhya*, vol. A25, no. 1, pp. 75–84, 1963.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[24] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–14, Jan. 1972.

[25] F. S. Hill, Jr. and M. A. Blanco, "Random geometric series and intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 326–335, May 1973.

[26] J. Ziv, "The Capacity of the General Time-Discrete Channel with Finite Alphabet," *Information and Control*, vol. 14, pp. 233–251, 1969.