

Information Theory Tutorial

Communication over Channels with memory

Chi Zhang
Department of Electrical Engineering
University of Notre Dame

A general capacity formula $C = \sup_X I(X; Y)$, which is correct for arbitrary single-user channels without feedback, is introduced in this tutorial. This new capacity formula is obtained by using a general channel model without any assumptions of the channel, introducing the notion of inf/sup-information/entropy rates, and finding a new tight converse bound on the error probability. In Section 1 we will show the error of the conventional channel capacity formula. Then in section 2, we will give some definitions used in the proofs of the capacity formula. Section 3 is devoted to the direct coding theorem. Section 4 will focus on the new converse bound and the general capacity formula. We will see some useful results of the new converse bound in section 5. In section 6, we will see how feedback can increase the channel capacity when the channel has memory.

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 $\exp(nR)$),

$$C = \max_X I(X; Y), \quad (1)$$

holds for *memoryless* channels. If the channel has memory, then (1) generalizes to the familiar limiting expression

$$C = \lim_{x \rightarrow 0} \sup_x \frac{1}{n} I(X^n; Y^n). \quad (2)$$

However, the capacity formula (2) does not hold in full generality. Let's see an example.

Example: 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. Obviously, the capacity of this channel is equal to 0. However the right-hand side of (2) is equal to 1/2 bit/channel use. Another counter example can be found at Nedomo [6]

The validity of (2) was proved for some specific class of channels, like information stable channels. Researchers also try to get a capacity formula for information unstable channels; however, all these approaches rely on some assumptions of the channel. There is a call for a completely general formula for channel capacity. The desired channel capacity is expressed in terms of the probabilistic description of the channel, which doesn't require any assumption of the channel, such as memorylessness, information stability, stationarity, causality, etc. Such a formula is found in "A General Formula for Channel Capacity" [2], which will be the main topic of this tutorial.

Before introducing the new channel capacity formula, we need several definitions.

Definition 1: Liminf in Probability

If A_n is a sequence of random variables, its liminf in probability is the supremum of all the reals α for which $P[A_n \geq \alpha] \rightarrow 0$ as $n \rightarrow \infty$.

Limsup in Probability

If A_n is a sequence of random variables, its limsup in probability is the infimum of all the reals β for which $P[A_n \geq \beta] \rightarrow 0$ as $n \rightarrow \infty$.

Definition 2: A channel W with input and output alphabets, A and B , respectively, is a sequence of conditional distributions $W = \{W^n(y^n|x^n) = P_{Y^n|X^n}(y^n|x^n); (x^n, y^n) \in A^n \times B^n\}_{n=1}^{\infty}$.

Remarks:

The channel definition in this paper is very general, without placing any restriction on the channel. It's also a reasonable channel model since it captures the physical situation to be modeled where block codewords are transmitted through the channel. For the convenience sake, in the proofs of this general channel capacity formula, we will assume that the input and output alphabets are finite; however, the proofs don't depend on this assumption.

Definition 3: Given a joint distribution $P_{Y^n|X^n}(y^n|x^n) = P_{X^n}(x^n)W^n(y^n|x^n)$, the information density is the function defined on $A^n \times B^n$:

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

Remarks:

The distribution of the random variable $(1/n)i_{X^n W^n}(X^n, Y^n)$ where X^n and Y^n have joint distribution $P_{X^n Y^n}$ will be referred to as the **information spectrum**. The expected value of the information spectrum is the normalized mutual information $(1/n)I(X^n; Y^n)$.

Definition 4: Inf-Information Rate

The liminf in probability of the sequence of random variables $(1/n)i_{X^n W^n}(X^n; Y^n)$ will be referred to as the the inf-information rate of the pair (X, Y) and will be denoted as $\underline{I}(X; Y)$.

Sup-Information Rate

The limsup in probability of the sequence of random variables $(1/n)i_{X^n W^n}(X^n; Y^n)$ will be referred to as the the sup-information rate of the pair (X, Y) and will be denoted as $\bar{I}(X; Y)$.

Remarks:

The introduction of inf/sup-information/entropy rates enables us to deal with nonergodic/nonstationary sources.

Using these definitions, the general capacity formula is

$$C = \sup_X \underline{I}(X; Y). \quad (3)$$

In (3), X denotes an input process in the form of a sequence of finite-dimensional distributions $X = \{X^n = (X_1^n, \dots, X_n^n)\}_{n=1}^\infty$. We denote by $Y = \{Y^n = (Y_1^n, \dots, Y_n^n)\}_{n=1}^\infty$ the corresponding output sequence of finite-dimensional distributions induced by X via the channel $W = \{W^n = P_{Y^n|X^n} : A^n \rightarrow B^n\}_{n=1}^\infty$, 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.

Before the proof of this new capacity formula, we will show some important properties of inf-information rate.

Theorem 1: 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 $\underline{I}(X; Y) = \lim_{n \rightarrow \infty} (1/n)I(X^n; Y^n)$
- h) $\underline{I}(X; Y) \leq \liminf_{n \rightarrow \infty} (1/n)I(X^n; Y^n)$.
- i) (**Data Processing Theorem**) $\underline{I}(X_1; X_3) \leq \underline{I}(X_1; X_2)$, if for every n , X_1 and X_3 are conditionally independent given X_2 .

Definition 5:

(1) The inf-divergence rate $\underline{D}(X \| Y)$ is defined for two arbitrary processes U and V as the liminf in probability of the sequence of log-likelihood ratios $\frac{1}{n} \log \frac{P_{U^n}(U^n)}{P_{V^n}(V^n)}$

(2) The sup-entropy rate $\overline{H}(Y)$ and inf-entropy $\underline{H}(Y)$ are defined as the limsup and liminf, respectively, in probability of the normalized entropy density $\frac{1}{n} \log \frac{1}{P_{Y^n}(y^n)}$.

Similarly, the conditional sup-entropy rate $\overline{H}(Y|X)$ is the limsup in probability (according to $P_{X^n Y^n}$) of $\frac{1}{n} \log \frac{1}{P_{Y^n|X^n}(y^n|x^n)}$.

Proof: The proofs can be found at [2]. Here we will try to interpret the above results.

Remarks:

(1) Many of the familiar properties satisfied by mutual information turn out to be inherited by the inf-information rate.

(2) The nonnegativity of the inf-divergence rate $\underline{D}(X \| Y)$ plays a key role in the proof of these properties,

just like its counterpart, the nonnegativity of divergence, in the proof of many of the mutual information properties.

(3)The proof of (e) can be found at [3]. That paper tells us that the minimum achievable source coding rate for any finite-alphabet source $X = X_{n=1}^{\infty}$ is equal to its sup-entropy rate $\overline{H}(X)$, defined as the limsup in probability of $(1/n) \log 1/P_{X^n}(X^n)$.

(4)The proof of (g) can also be found at [3]. When (X,Y) satisfies this property, the input-output pair (X,Y) is called information stable.

III. DIRECT CODING THEOREM:

Definition 6: An (n, M, ϵ) code has block length n , M codewords, and error probability 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 $C = \lim_{\epsilon \rightarrow 0} C_\epsilon$

Theorem 2 (Feinstein's lemma): 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

$$\epsilon \leq P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n} \log M + \gamma\right] + \exp(-\gamma n) \quad (4)$$

The direct part of the coding theorem follows Feinstein's lemma and the definitions of capacity and inf-information rate.

Theorem 3:

$$C \geq \sup_X \underline{I}(X; Y). \quad (5)$$

Proof:

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

$$\underline{I}(X; Y) - \delta < \frac{\log M}{n} < \underline{I}(X; Y) - \frac{\delta}{2} \quad (6)$$

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

$$P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n} \log M + \delta/4\right] \leq P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \underline{I}(X; Y) - \delta/4\right] \leq \frac{\epsilon}{2} \quad (7)$$

where the second inequality holds for all sufficiently large n because of the definition of $\underline{I}(X; Y)$.⁵ In view of (7), Theorem 2 guarantees the existence of the desired codes. □

IV. CONVERSE CODING THEOREM:

In this section, we will introduce a new converse, which is the main result of [2]. It's tight for every channel, and it's obtained without recourse to the Fano inequality. We will also compare this new converser bound with the conventional Fano inequality, which is short of providing a tight converse for some channels.

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

$$\epsilon \geq P\left[\frac{1}{n}i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n} \log M - \gamma\right] - \exp(-\gamma n) \quad (8)$$

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 (8) is equal to the set of "atypical" input-output pairs

$$L = \{(a^n, b^n) \in A^n \times B^n : P_{X^n|Y^n}(a^n|b^n) \leq \beta\} \quad (9)$$

This is because the information density can be written as

$$i_{X^n W^n}(a^n; b^n) = \log \frac{P_{X^n|Y^n}(a^n|b^n)}{P_{X^n}(a^n)} \quad (10)$$

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

We need to show that

$$P_{X^n Y^n}[L] \leq \epsilon + \beta \quad (11)$$

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

$$B_i = \{b^n \in B^n : P_{X^n|Y^n}(c_i|b^n) \leq \beta\} \quad (12)$$

We can write

$$P_{X^n Y^n}[L] = \sum_{i=1}^M P_{X^n Y^n}[(c_i, B_i)] \quad (13)$$

$$= \sum_{i=1}^M P_{X^n Y^n}[(c_i, B_i \cap D_i^c)] + \sum_{i=1}^M P_{X^n Y^n}[(c_i, B_i \cap D_i)] \quad (14)$$

$$\leq \sum_{i=1}^M \frac{1}{M} W^n(D_i^c|c_i) + \beta P_{Y^n}\left[\bigcup_{i=1}^M D_i\right] \quad (15)$$

$$\leq \epsilon + \beta \quad (16)$$

where the second inequality is due to (12) and the disjointness of the decoding sets.

Remark:

(1) This converse shows a dual of Feinstein's result: the average error probability of any code is essentially lower-bounded by the cumulative distribution function of the input-output information density evaluated at the code rate.

(2) Theorem 4 gives a family (parameterized by γ) of lower bounds on the error probability. To obtain the best bound, we simply maximize the right-hand side of (8) over γ .

(3) A weaker bound

Using the Fano inequality, every (n, M, ϵ) code satisfies

$$\log M \leq \frac{1}{1 - \epsilon} [I(X^n; Y^n) + h(\epsilon)]; \quad (17)$$

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.

Using (17), it is evident that if $R \geq 0$ is ϵ -achievable, then for every $\delta > 0$.

$$R - \delta < \frac{1}{1 - \epsilon} \left[\frac{1}{n} \sup_{X^n} I(X^n; Y^n) + \frac{h(\epsilon)}{n} \right] \quad (18)$$

which, in turn, implies

$$R \leq \frac{1}{1 - \epsilon} \liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{X^n} I(X^n; Y^n). \quad (19)$$

Thus, the general converse in (2) follows by letting $\epsilon \rightarrow 0$

This weak bound is unable to provide the desired tight converse because it depends on the channel through the input-output mutual information (expectation of information density) achieved by the code. However, the new converse bound only depends on the distribution of the information density achieved by the code, rather than on just its expectation.

Theorem 5:

$$C \leq \sup_X \underline{I}(X; Y). \quad (20)$$

Proof:

Using the concept of the information spectrum [3], Theorem 4 tells us 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. With this insight, the proof in [2] use a contradiction to prove the converse part of the channel coding theorem.

□

We can use the results above 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_ϵ ,

$$C_\epsilon \leq \sup_X \sup R : F_X(R) \leq \epsilon \quad (21)$$

$$C_\epsilon \geq \sup_X \sup R : F_X(R) < \epsilon \quad (22)$$

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

$$F_X(R) = \lim_n \sup P[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \leq R]. \quad (23)$$

The bounds (21) (22) hold with equality, except possibly at the points of discontinuity of C_ϵ , of which there are, at most, countably many.

VI. GAUSSIAN FEEDBACK CAPACITY

Pinsker and Ebert showed that feedback at most doubles the capacity of a non-white Gaussian channel; a simple proof can be found in Cover and Pombra [4].

Lemma 1: For A,B nonnegative definite matrices and $0 \leq \lambda \leq 1$

$$|\lambda A + (1 - \lambda)B| \geq |A|^\lambda |B|^{1-\lambda} \quad (24)$$

Lemma 2: If X^n and Z^n are causally related, then

$$h(X - Z) \geq h(Z) \quad (25)$$

and

$$|K_{X-Z}| \geq |K_Z| \quad (26)$$

Remark:

The random vector X^n is causally related to Z^n if $f(x^n, Z^n) = f(z^n) \prod_{i=1}^n f(x_i | x^{i-1}, z^{i-1})$. Note that feedback codes necessarily yield causally related (X^n, Z^n) .

Theorem 7: $C_{n,FB} \leq 2C_n$.

$$\frac{1}{2n} \log \frac{|K_{X+Z}|}{K_Z} = \frac{1}{2n} \log \frac{|\frac{1}{2}K_{X+Z} + \frac{1}{2}K_{X-Z}|}{K_Z} \quad (27)$$

$$\geq \frac{1}{2n} \log \frac{|K_{X+Z}|^{1/2} |K_{X-Z}|^{1/2}}{K_Z} \quad (28)$$

$$\geq \frac{1}{2n} \log \frac{|K_{X+Z}|^{1/2} |K_Z|^{1/2}}{K_Z} \quad (29)$$

$$= \frac{1}{2} \frac{1}{2n} \log \frac{|K_{X+Z}|}{K_Z} \quad (30)$$

By maximizing each side, we have

$$\frac{1}{2}C_{n,FB} \leq C_{nd} \quad (31)$$

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] S. Verdu and T. S. Han, A General Formula for Channel Capacity, IEEE Trans. Inform. Theory, July 1994.
- [3] T. S. Han and S. Verdu, Approximation theory of output statistics, IEEE Trans. Inform. Theory, vol. 39, pp. 752-772, May 1993.
- [4] T. M. Cover and S. Pombra, Gaussian feedback capacity, IEEE Trans. Inform. Theory, vol. 35, pp. 37C43, Jan. 1989.
- [5] A. Feinstein, A new basic theorem of information theory, IRE Trans. Inform. Theory, vol. IT-4, pp. 2-22, 1954.
- [6] I. Nedom, The capacity of a discrete channel, in Proc. 1st Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes, Prague, 1957, pp. 143-181.