Generalized source coding theorems and hypothesis testing: Part I — Information measures

Po-Ning Chen \textsuperscript{a} & Fady Alajaji \textsuperscript{b}

\textsuperscript{a} Dept. of Communications Engineering, National Chiao Tung University, Hsin Chu, Taiwan 300, R.O.C.
\textsuperscript{b} Dept. of Mathematics and Statistics, Queen's University, Kingston, Ontario, K7L 3N6, Canada

Published online: 03 Mar 2011.


To link to this article: http://dx.doi.org/10.1080/02533839.1998.9670392
GENERALIZED SOURCE CODING THEOREMS AND HYPOTHESIS TESTING: PART I -- INFORMATION MEASURES

Po-Ning Chen*
Dept. of Communications Engineering
National Chiao Tung University
Hsin Chu, Taiwan 300, R.O.C.

Fady Alajaji
Dept. of Mathematics and Statistics
Queen’s University
Kingston, Ontario K7L 3N6, Canada

Key Words: information theory, entropy, mutual information, divergence, e-capacity.

ABSTRACT

Expressions for ε-entropy rate, ε-mutual information rate and ε-divergence rate are introduced. These quantities, which consist of the quantiles of the asymptotic information spectra, generalize the inf/sup-entropy/information/divergence rates of Han and Verdú. The algebraic properties of these information measures are rigorously analyzed, and examples illustrating their use in the computation of the e-capacity are presented. In Part II of this work, these measures are employed to prove general source coding theorems for block codes, and the general formula of the Neyman-Pearson hypothesis testing type-II error exponent subject to upper bounds on the type-I error probability.

I. INTRODUCTION AND MOTIVATION

Entropy, divergence and mutual information are without a doubt the most important information theoretic quantities. They constitute the fundamental measures upon which information theory is founded. Given a discrete random variable X with distribution \( P_X \), its entropy is defined by

\[
H(X) = -\sum x P(x) \log_2 P(x) = E_{P_X}[- \log P(x)].
\]

As for the mutual information \( I(X;Y) \) between random variables X and Y, it represents the average amount of information that Y contains about X. It is defined as the divergence between the joint distribution \( P_{XY} \) and the product distribution \( P_X P_Y \):

\[
I(X;Y) = D(P_{XY} \| P_X P_Y) = E_{P_{XY}}[\log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}].
\]

More generally, consider an input process X defined by a sequence of finite dimensional distributions \([11]\): \( X^n = (X_1^{(n)}, ..., X_n^{(n)})^{n}_{n=1} \). Let \( Y^n = (Y_1^{(n)}, ..., Y_n^{(n)})^{n}_{n=1} \) be the corresponding output process induced by X via the channel \( W^n = (W_1^{(n)}, ..., W_n^{(n)})^{n}_{n=1} \), which is an arbitrary sequence of n-dimensional conditional distributions from \( X^n \) to \( Y^n \), where \( X \) and \( Y \) are the input and output alphabets respectively. The entropy rate for the source \( X^n \) is defined by[2], [7]

As for the mutual information \( I(X;Y) \) between random variables X and Y, it represents the average amount of information that Y contains about X. It is defined as the divergence between the joint distribution \( P_{XY} \) and the product distribution \( P_X P_Y \):

\[
I(X;Y) = D(P_{XY} \| P_X P_Y) = E_{P_{XY}}[\log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}].
\]
assuming the limit exists. Similarly the expressions for the divergence and mutual information rates are given by

\[ D(X \mid \hat{X}) = \lim_{n \to \infty} \frac{1}{n} E[\log \frac{P_{\hat{X}}(X^n)}{P_X^n}] \]

and

\[ I(X;Y) = \lim_{n \to \infty} \frac{1}{n} E[\log \frac{P_{X,Y^n}(X^n,Y^n)}{P_X^n P_{Y^n}}] , \]

respectively.

The above quantities have an operational significance established via Shannon’s coding theorems when the stochastic systems under consideration satisfy certain regularity conditions (such as stationarity and ergodicity, or information stability) \[9\], \[11\]. However, in more complicated situations such as when the systems are non-stationary (with time-varying statistics), these information rates are no longer valid and lose their operational significance. This results in the need to establish new information measures which appropriately characterize the operational limits of arbitrary stochastic systems.

This is achieved in \[10\] and \[11\] where Han and Verdú introduce the notions of \textit{inf-sup-entropy/information rates} and illustrate the key role these information measures play in proving a general lossless (block) source coding theorem and a general channel coding theorem. More specifically, they demonstrate that for an arbitrary finite-alphabet source \( X \), the expression for the minimum achievable (block) source coding rate is given by the \textit{sup-entropy rate} \( H(X) \), defined as the \( \limsup \) in probability of \((1/n) \log 1/P_{X^n}(X^n)\) \[10\]. They also establish in \[11\] the formulas of the \( \epsilon \)-capacity \( C_\epsilon \) and capacity\(^1\) \( C \) of arbitrary single-user channels without feedback (not necessarily information stable, stationary, ergodic, etc.). More specifically, they show that

\[ \sup_X \sup_{\epsilon \leq 1} \frac{1}{R} \{ F_X(R) \leq \epsilon \} \leq C_\epsilon \leq \sup_X \sup_{\epsilon \leq 1} \frac{1}{R} \{ F_X(R) \leq \epsilon \} , \]

and

\[ C = \sup_X I(X;Y) , \]

where

\[ F_X(R) = \lim_{n \to \infty} \Pr \left\{ \frac{1}{n} \log P_{X^n}(X^n) \leq R \right\} , \]

\((1/n) i_{X^n}\gamma\mu(X^n; Y^n)\) is the sequence of normalized information densities defined by

\[ i_{X^n}\gamma\mu(X^n; Y^n) = \log \frac{P_{X^n\gamma\mu}(X^n,Y^n) P_{Y^n}}{P_X^n} , \]

and \( I(X;Y) \) is \textit{inf-information rate} between \( X \) and \( Y \), which is defined as the \( \liminf \) in probability of \((1/n) i_{X^n}\mu\gamma(X^n; Y^n)\).

By adopting the same technique as in \[10\] (also in \[11\]), general expressions for the capacity of single-user channels with feedback and for Neyman-Pearson type-II error exponents are derived in \[4\] and \[6\], respectively. Furthermore, an application of the type-II error exponent formula to the non-feedback and feedback channel reliability functions is demonstrated in \[6\] and \[5\].

The above \textit{inf-sup-entropy/information rates} are expressed in terms of the \textit{liminf/limsup} of the normalized entropy/information densities. The \textit{lim inf} in probability of a sequence of random variables is defined as follows \[10\]: if \( A_n \) is a sequence of random variables, then its \textit{lim inf} in probability is the largest extended real number \( U \) such that for all \( \xi > 0 \),

\[ \lim_{n \to \infty} \Pr [A_n \leq U - \xi] = 0 . \tag{1} \]

Similarly, its \textit{lim sup} in probability is the smallest extended real number \( U \) such that for all \( \xi > 0 \),

\[ \lim_{n \to \infty} \Pr [A_n \geq U + \xi] = 0 . \tag{2} \]

Note that these two quantities are always defined; if they are equal, then the sequence of random variables converges in probability to a constant.

It is straightforward to deduce that Eqs. (1) and (2) are respectively equivalent to

\[ \liminf_{n \to \infty} \Pr [A_n \leq U - \xi] = \limsup_{n \to \infty} \Pr [A_n \leq U - \xi] = 0 . \tag{3} \]

and

\[ \liminf_{n \to \infty} \Pr [A_n \geq U + \xi] = \limsup_{n \to \infty} \Pr [A_n \geq U + \xi] = 0 . \tag{4} \]

\(^1\text{Definition} \ ([8], [11]): \text{Given} 0 < \epsilon < 1, \text{an} (n, M, \epsilon) \text{code for the channel} W \text{has blocklength} n, M \text{codewords and average (decoding) error probability not larger than} \epsilon. \text{A non-negative number} R \text{is an} \epsilon\text{-achievable rate} \text{if for every} \delta > 0, \text{there exist, for all} n \text{sufficiently large,} (n, M, \epsilon) \text{codes with rate} (1/n) \log M > R - \delta. \text{The supremum of all} \epsilon\text{-achievable rates is called the} \epsilon\text{-capacity,} C_\epsilon. \text{The capacity} C \text{is the supremum of rates that are} \epsilon\text{-achievable for all} 0 < \epsilon < 1 \text{and hence} C = \lim_{\epsilon \to 0} C_\epsilon . \]

In other words, \( C_\epsilon \) is the largest rate at which information can be conveyed over the channel such that the probability of a decoding error is below a fixed threshold \( \epsilon \), for sufficiently large blocklengths. Furthermore, \( C \) represents the largest rate at which information can be transmitted over the channel with asymptotically \textit{vanishing} error probability.
We can observe however that there might exist cases of interest where only the liminf of the probabilities in (3) and (4) are equal to zero; while the limsup may not vanish. There are also other cases where both the liminf and limsup in (3)-(4) do not vanish; but they are upper bounded by a prescribed threshold. Furthermore, there are situations where the interval \([U, \bar{U}]\) does not contain only one point for e.g., when \(A_n\) converges in distribution to another random variable. Hence, those points within the interval \([U, \bar{U}]\) might possess a Shannon-theoretic operational meaning when for example \(A_n\) consists of the normalized entropy density of a given source.

The above remarks constitute the motivation for this work in which we generalize Han and Verdú's information rates and prove general data compression and hypothesis testing theorems that are the counterparts of their \(\epsilon\)-capacity channel coding theorem [11].

In Part I, we propose generalized versions of the inf/sup-entropy/information/divergence rates. We analyze in detail the algebraic properties of these information measures, and we illustrate their use in the computation of the \(\epsilon\)-capacity of arbitrary additive-noise channels. In Part II of this paper [3], we utilize these quantities to establish general source coding theorems for arbitrary finite-alphabet sources, and the general expression of the Neyman-Pearson type-II error exponent.

II. GENERALIZED INFORMATION MEASURES

Definition 1. (Inf/sup-spectrum)

If \(\{A_n\}_{n=1}^{\infty}\) is a sequence of random variables, then its inf-spectrum \(\underline{u}(\cdot)\) and its sup-spectrum \(\overline{u}(\cdot)\) are defined by

\[
\underline{u}(\theta) = \lim_{n \to \infty} \inf \Pr\{A_n \leq \theta\},
\]

and

\[
\overline{u}(\theta) = \lim_{n \to \infty} \sup \Pr\{A_n \leq \theta\}.
\]

In other words, \(\underline{u}(\cdot)\) and \(\overline{u}(\cdot)\) are respectively the liminf and the limsup of the cumulative distribution function (CDF) of \(A_n\). Note that by definition, the CDF of \(A_n - \Pr\{A_n \leq \theta\}\) is non-decreasing and right-continuous. However, for \(\underline{u}(\cdot)\) and \(\overline{u}(\cdot)\), only the non-decreasing property remains.\(^2\)

Definition 2. (Quantile of inf/sup-spectrum)

For any \(0 \leq \delta < 1\), the quantiles \(\underline{U}_\delta\) and \(\overline{U}_\delta\) of the sup-spectrum and the inf-spectrum are defined by\(^3\)

\[
\underline{U}_\delta = \begin{cases} 
-\infty & \text{if } \{\theta : \underline{u}(\theta) \leq \delta\} = \emptyset, \\
\sup\{\theta : \underline{u}(\theta) \leq \delta\}, & \text{otherwise},
\end{cases}
\]

and

\[
\overline{U}_\delta = \begin{cases} 
-\infty & \text{if } \{\theta : \overline{u}(\theta) \leq \delta\} = \emptyset, \\
\sup\{\theta : \overline{u}(\theta) \leq \delta\}, & \text{otherwise},
\end{cases}
\]

respectively. It follows from the above definitions that \(\underline{U}_\delta\) and \(\overline{U}_\delta\) are right-continuous and non-decreasing in \(\delta\).

Note that the liminf in probability \(U\) and the limsup in probability \(\bar{U}\) of \(A_n\) satisfy

\[
U = U_0
\]

and

\[
\overline{U} = \overline{U}_1^-,
\]

respectively, where the superscript "−" denotes a strict inequality in the definition of \(\overline{U}_1^-\); i.e.,

\[
\overline{U}_\delta = \sup\{\theta : \overline{u}(\theta) < \delta\}.
\]

Note also that

\[
U \leq \underline{U}_\delta \leq \overline{U}_\delta \leq \overline{U}.
\]

Remark that \(\underline{U}_\delta\) and \(\overline{U}_\delta\) always exist. Furthermore, if \(U = U_\delta\ \forall \delta \in [0, 1]\), then the sequence of random variables \(A_n\) converges in distribution to a random variable \(A\), provided the distribution sequence of \(A_n\) is tight.

For a better understanding of the quantities defined above, we depict them in Fig. 1.

In the above definitions, if we let the random variable \(A_n\) equal the normalized entropy density of an arbitrary source \(X\), we obtain two generalized entropy measures for \(X\): the \(\delta\)-inf-entropy-rate \(H_\delta(X)\)

\(^2\) It is pertinent to also point out that even if we do not require right-continuity as a fundamental property of a CDF, the spectra \(u(\cdot)\) and \(\overline{u}(\cdot)\) are not necessarily legitimate CDFs of (conventional real-valued) random variables since there might exist cases where the "probability mass escapes to infinity" (cf. [1, page 346]). A necessary and sufficient condition for \(u(\cdot)\) and \(\overline{u}(\cdot)\) to be conventional CDFs (without requiring right-continuity) is that the sequence of distribution functions of \(A_n\) be tight [1, page 346]. Tightness is actually guaranteed if the alphabet of \(A_n\) is finite.

\(^3\) Note that the usual definition of the quantile function \(\phi(\delta)\) of a non-decreasing function \(F(\cdot)\) is slightly different from our definition [1, page 190]: \(\phi(\delta) = \sup\{\theta : F(\theta) < \delta\}\). Remark that if \(F(\cdot)\) is strictly increasing, then the quantile is nothing but the inverse of \(F(\cdot)\): \(\phi(\delta) = F^{-1}(\delta)\).
and the $\delta$-sup-entropy-rate $\overline{H}_\delta(X)$ as described in Table 1. Note that the inf-entropy-rate $H(X)$ and the sup-entropy-rate $\underline{H}(X)$ introduced in [10] are special cases of the $\delta$-inf/sup-entropy rate measures:

$$H(X) = H_0(X), \quad \text{and} \quad \overline{H}_\delta(X) = \underline{H}_{-\delta}(X).$$

Analogously, for an arbitrary channel $W = P_{Y/X}$ with input $X$ and output $Y$ (or respectively for two observations $X$ and $X'$), if we replace $A_n$ by the normalized information density (resp. by the normalized log-likelihood ratio), we get the $\delta$-inf/sup-information rates (resp. $\delta$-inf/sup-divergences rates) as shown in Table 1.

The algebraic properties of these newly defined information measures are investigated in the next section.

III. PROPERTIES OF THE GENERALIZED INFORMATION MEASURES

Lemma 1.

Consider two arbitrary random sequences, \( \{A_n\}_{n=1}^\infty \) and \( \{B_n\}_{n=1}^\infty \). Let $\underline{u}(\cdot)$ and $\underline{u}(\cdot)$ denote respectively the sup-spectrum and inf-spectrum of \( \{A_n\}_{n=1}^\infty \). Similarly, let $\overline{u}(\cdot)$ and $\overline{u}(\cdot)$ denote respectively the sup-spectrum and inf-spectrum of \( \{B_n\}_{n=1}^\infty \). Define $\underline{U}_\delta = \sup \{ \theta : \underline{u}(\theta) \leq 0 \}$, $\overline{U}_\delta = \sup \{ \theta : \overline{u}(\theta) \leq 0 \}$, $\underline{Y}_\delta = \sup \{ \theta : \underline{u}_\delta(\theta) \leq 0 \}$, $\overline{V}_\delta = \sup \{ \theta : \overline{u}_\delta(\theta) \leq 0 \}$, then the following statements hold.

1. $\underline{U}_\delta$ and $\overline{U}_\delta$ are both non-decreasing functions of $\delta \in [0,1]$.
2. For $\delta \geq 0$, $\gamma \geq 0$, and $1 \geq \delta + \gamma$,

\[ (\underline{U} + \overline{V})_{\delta + \gamma} \leq \underline{U}_\delta + \overline{V}_\gamma. \]

3. For $\delta \geq 0$, $\gamma \geq 0$, and $1 > \delta + \gamma$,

\[ (\underline{U} + \overline{V}) \leq \underline{U}_\delta + \overline{V}_{(1-\gamma)}, \]

and

$$\underline{U}_\delta \geq \overline{V}_\delta, \quad \text{and} \quad \overline{U}_\delta \geq \underline{V}_\delta.$$
Table 1. Generalized information measures where $\delta \in [0,1]$.

<table>
<thead>
<tr>
<th>Entropy Measures</th>
<th>System</th>
<th>(A_n): Norm. Entropy Density</th>
<th>(\frac{1}{n} h_{X^n}(X^n) \leq -(1/n) \log P_{X^n}(X^n))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy Sup-Spectrum</td>
<td>(\bar{h}<em>X(\theta) = \lim \sup \Pr {(1/n) h</em>{X^n}(X^n) \leq \theta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entropy Inf-Spectrum</td>
<td>(\underline{h}<em>X(\theta) = \lim \inf \Pr {(1/n) h</em>{X^n}(X^n) \leq \theta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\delta)-Inf-Entropy Rate</td>
<td>(H_\delta(X) = \sup {\theta: \underline{h}_X(\theta) \leq \delta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\delta)-Sup-Entropy Rate</td>
<td>(\overline{H}_\delta(X) = \sup {\theta: \bar{h}_X(\theta) \leq \delta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sup-Entropy Rate</td>
<td>(\overline{H}(X) = \overline{H}_1(X))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inf-Entropy Rate</td>
<td>(H(\theta) = H_0(X))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mutual Information Measures</th>
<th>System</th>
<th>(A_n): Norm. Information Density</th>
<th>(\frac{1}{n} i_{X^n,Y^n}(X^n;Y^n) \Delta = \frac{1}{n} \log \frac{dP_{X^n,Y^n}}{d(P_{X^n} \times P_{Y^n})}(X^n,Y^n))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Sup-Spectrum</td>
<td>(\bar{i}<em>{X,Y}(\theta) = \lim \sup \Pr {(1/n) i</em>{X^n,Y^n}(X^n;Y^n) \leq \theta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Inf-Spectrum</td>
<td>(\underline{i}<em>{X,Y}(\theta) = \lim \inf \Pr {(1/n) i</em>{X^n,Y^n}(X^n;Y^n) \leq \theta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\delta)-Inf-Information Rate</td>
<td>(I_\theta(X;Y) = \sup {\theta: \underline{i}_{X,Y}(\theta) \leq \delta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\delta)-Sup-Information Rate</td>
<td>(\overline{I}<em>\theta(X;Y) = \sup {\theta: \bar{i}</em>{X,Y}(\theta) \leq \delta})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sup-Information Rate</td>
<td>(\overline{I}(X;Y) = \overline{I}_1(X;Y))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inf-Information Rate</td>
<td>(I(X;Y) = I_0(X;Y))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Divergence Measures | System | \(A_n\): Norm Log-Likelihood Ratio | \(\frac{1}{n} d_{X^n}(X^n \mid \tilde{X}^n) \Delta = \frac{1}{n} \log \frac{dP_{X^n|\tilde{X}^n}}{dP_{X^n}}(X^n)\) |
|---------------------|--------|-------------------------------|--------------------------------------------------|
| Divergence Sup-Spectrum | \(\overline{d}_{X^n} \{ \tilde{X}^n(\theta) = \lim \sup \Pr \{(1/n) d_{X^n}(X^n \mid \tilde{X}^n) \leq \theta\}\) |
| Divergence Inf-Spectrum | \(\underline{d}_{X^n} \{ \tilde{X}^n(\theta) = \lim \inf \Pr \{(1/n) d_{X^n}(X^n \mid \tilde{X}^n) \leq \theta\}\) |
| \(\delta\)-Inf-Divergence Rate | \(D_\delta(X \mid \tilde{X}) = \sup \{\theta: \overline{d}_{X^n} \{ \tilde{X}^n(\theta) \leq \delta\}\) |
| \(\delta\)-Sup-Divergence Rate | \(\overline{D}_\delta(X \mid \tilde{X}) = \sup \{\theta: \underline{d}_{X^n} \{ \tilde{X}^n(\theta) \leq \delta\}\) |
| Sup-Divergence Rate | \(\overline{D}(X \mid \tilde{X}) = \overline{D}_1(X \mid \tilde{X})\) |
| Inf-Divergence Rate | \(D(X \mid \tilde{X}) = D_0(X \mid \tilde{X})\) |

\[(U+\gamma) \leq U_{\delta+\gamma}(-V)\gamma\]

(Note that the cases \(\delta+\gamma=1\) or \(\gamma=1\) are not allowed here because they result in \(U_{1}=V_{1}=\infty\), and the subtraction of two infinite terms is undefined. That is why the condition for property 2, \(1 \leq \delta+\gamma\) is replaced by \(1 > \delta+\gamma\) in property 3.)

The proof is completed by showing that

\[\overline{ \left( -V \right) } \Delta = \overline{V}(0 \left( -V \right) ).\] (9)

By definition,

\[\overline{v}(\theta) = \lim \sup \Pr \{B_n \leq \theta\}\]
So \( v(\theta^* + \epsilon) = 1 - v(\theta) \). Then

\[
\gamma = \liminf_{n \to \infty} \Pr \{ (A_n + B_n) < \theta \} = 1 - v(\theta^*).
\]

and

\[
\gamma = \limsup_{n \to \infty} \Pr \{ (A_n + B_n) > \theta \} = 1 - v(\theta^*).
\]

where the inequality follows from \( v(\theta) \geq v(\theta^*) \). Finally,

\[
\gamma = \liminf_{n \to \infty} \Pr \{ (A_n + B_n) < \theta \} = 1 - v(\theta^*).
\]

Using (9), we have the desired result.

If we take \( \delta_\gamma = \gamma = 0 \) in (5) and (7), we obtain

\[
(U + V) = \mathcal{L}(U + V), \quad \text{and} \quad (U + V) \leq \mathcal{L}(U + V),
\]

which mean that the liminf in probability of a sequence of random variables \( A_n + B_n \) is upper [resp. lower] bounded by the liminf in probability of \( A_n \) plus the limsup [resp. liminf] in probability of \( B_n \). This fact is used in [11] to show that

\[
\mathcal{H}(Y) - \mathcal{H}(Y \mid X) \leq \mathcal{L}(X;Y) \leq \mathcal{H}(Y) - \mathcal{H}(Y \mid X),
\]

which is a special case of property 3 in Lemma 2.

The next lemmas will show some of the analogous properties of the generalized information measures.

Lemma 2.

For \( \delta, \gamma, \delta + \gamma \in [0,1) \), the following statements hold.

1. \( \mathcal{H}_\delta(X) \geq \mathcal{L}_\delta(X) \geq \mathcal{I}_\delta(X) \geq \mathcal{H}_\delta(X) \geq \mathcal{H}(X) \).

2. \( \mathcal{H}_\delta(X;Y) \geq \mathcal{I}_\delta(X;Y) \geq \mathcal{L}_\delta(X;Y) \).

3. \( \mathcal{H}_\delta(X;Y) \geq \mathcal{H}_\delta(Y \mid X) \).

4. \( \mathcal{L}_\delta(X;Y) \leq \mathcal{L}_\delta(Y \mid X) \).

5. \( \mathcal{I}_\delta(X;Y) \leq \mathcal{I}_\delta(Y \mid X) \).

\[ \mathcal{H}_\delta(X;Y) \leq \mathcal{H}_\delta(Y \mid X) \]

\[ \mathcal{L}_\delta(X;Y) \leq \mathcal{L}_\delta(Y \mid X) \]

where \( \delta \) is the minimum achievable (i.e., with asymptotically negligible probability of decoding error) fixed-length coding rate for \( X \) as seen in [3, Theorem 3.2] and [10].

Property 5 can be proved using the fact that

\[
\frac{1}{n} \log \Pr \left\{ X^n \mid Y^n \right\} = \frac{1}{n} \log \Pr \left\{ X^n \mid Y^n \right\}
\]

with this fact, (10) follows directly from (5), (11) and (12) follow from (6), (13) follows from (7), and (14) follows from (8).

Property 4 follows from the fact that \( \mathcal{H}_\delta(-) \) is non-decreasing in \( \delta \); \( \mathcal{H}_\delta(X) \leq \mathcal{H}_\delta(-X) = \mathcal{H}(X) \), and that \( \mathcal{H}(X) \) is the minimum achievable (i.e., with asymptotically negligible probability of decoding error) fixed-length coding rate for \( X \) as seen in [3, Theorem 3.2] and [10].

Property 5 can be proved using the fact that

\[
\frac{1}{n} \log \Pr \left\{ X^n \mid Y^n \right\} = \frac{1}{n} \log \Pr \left\{ X^n \mid Y^n \right\}
\]

By applying (5), and letting \( \gamma = 0 \), we obtain the desired result.
Lemma 3. (Data processing lemma)

Fix $\delta \in [0,1)$. Suppose that for every $n$, $X^n_1$ and $X^n_3$ are conditionally independent given $X^n_2$. Then

$$I_\delta(X^n_1;X^n_3) \leq I_\delta(X^n_1;X^n_2).$$

Proof:
By property 5, we get

$$I_\delta(X^n_1;X^n_3) \leq I_\delta(X^n_1;X^n_2),$$

where the equality holds because

$$\frac{1}{n} \log \frac{dP_{X^n_1|X^n_3}}{d(P_{X^n_1} \times P_{X^n_3})}(X^n_1, X^n_3) = \frac{1}{n} \log \frac{dP_{X^n_1}}{d(P_{X^n_1} \times P_{X^n_3})}(X^n_1, X^n_2).$$

Lemma 4. (Optimality of independent inputs)

Fix $\delta \in [0,1)$. Consider a finite alphabet, discrete memoryless channel – i.e., $P_{Y^n|x^n} = \prod_{i=1}^{n} P_{Y^n|x^n}$, for all $n$. For any input $X^n$ and its corresponding output $Y^n$, $I_\delta(X^n;Y^n) \leq I_\delta(X^n;Y^n) = I(X^n;Y^n)$, where $Y^n$ is the output due to $X^n$, which is an independent process with the same first order statistics as $X^n$, i.e., $P_{Y^n} = \prod_{i=1}^{n} P_{Y_i}$.

Proof:
First, we observe that

$$\frac{1}{n} \log \frac{dP_{Y^n|x^n}}{dP_{Y^n}}(X^n, Y^n) + \frac{1}{n} \log \frac{dP_{Y^n}}{dP_{Y^n}}(X^n, Y^n) = \frac{1}{n} \log \frac{dP_{Y^n|x^n}}{dP_{Y^n}}(X^n, Y^n).$$

In other words,

$$\frac{1}{n} \log \frac{dP_{Y^n|x^n}}{dP_{Y^n}}(X^n, Y^n) + \frac{1}{n} \log \frac{dP_{Y^n}}{dP_{Y^n}}(X^n, Y^n) = \frac{1}{n} \log \frac{dP_{Y^n|x^n}}{dP_{Y^n}}(X^n, Y^n).$$

By evaluating the above terms under $P_{X^n|Y^n}$ and letting

$$\overline{c}(\Theta) = \lim_{n \to \infty} \sup_{P_{X^n|Y^n}} \left( \frac{1}{n} \log \frac{dP_{X^n|Y^n}}{dP_{X^n|Y^n}}(X^n, Y^n) \leq \Theta \right),$$

we obtain from (5) (with $\gamma = 0$) that

$$Z_\delta(X^n;Y^n) \geq I_\delta(X^n;Y^n) + D(Y||Y^n) \geq I_\delta(X^n;Y^n),$$

since $D(Y||Y^n) \geq 0$ by property 1 of Lemma 2.

Note that the summable property of $(1/n)\log \left[ \frac{dP_{X^n|Y^n}}{d(P_{X^n|Y^n} \times P_{Y^n})}(X^n, Y^n) \right]$ (i.e., it is equal to $(1/n)\sum_{i=1}^{n} \log \left[ \frac{dP_{X^n|Y^n}}{d(P_{X^n|Y^n} \times P_{Y^n})}(X^n, Y^n) \right]$), the Chebyshev inequality and the finiteness of the channel alphabets imply

$$I(X^n;Y^n) = I_\delta(X^n;Y^n) \text{ and } Z(X^n;Y^n) = Z_\delta(X^n;Y^n).$$

It finally remains to show that

$$I(X^n;Y^n) \geq Z(X^n;Y^n),$$

which is proved in [11, Theorem 10].

IV. EXAMPLES FOR THE COMPUTATION OF \( \varepsilon \)-CAPACITY

In [11], Verdú and Han establish the general formulas for channel capacity and $\varepsilon$-capacity. In terms of the $\varepsilon$-inf-information rate, the expression of the $\varepsilon$-capacity becomes

$$\sup_{X} I_{\varepsilon}(X^n;Y^n) \leq C_{\varepsilon} \leq \sup_{X} I_{\varepsilon}(X^n;Y^n),$$

where $\varepsilon \in (0,1)$.

We now provide examples for the computation of $C_{\varepsilon}$. They are basically an extension of some of the examples provided in [11] for the computation of channel capacity. In this section, we assume that all the logarithms are in base 2.

Let the alphabet be binary $X = \{0, 1\}$, and let every output be given by

$$Y_i = X_i \oplus Z_i,$$

where $\oplus$ represents the addition operation modulo-2 and $Z$ is an arbitrary binary random process independent of $X$.

To compute the $\varepsilon$-capacity we use the results of property 3 in Lemma 2:

$$I_{\varepsilon}(X^n;Y^n) \geq H(Y^n) - H_{\varepsilon}(Y^n|X^n),$$

and

$$I_{\varepsilon}(X^n;Y^n) \leq \min(H_{\varepsilon}(Y^n|X^n), H_{\varepsilon}(Y^n|X^n), H_{\varepsilon}(Y^n|X^n), H_{\varepsilon}(Y^n|X^n)).$$

(16)
where $\varepsilon \geq 0$, $\gamma \geq 0$ and $1 > \varepsilon + \gamma$. The lower bound in (15) follows directly from (13) (by taking $\delta = 0$ and $\gamma = \varepsilon$). The upper bounds in (16) follow from (10) and (11) respectively.

\[
C_e \leq \sup_x \{ H_{e+\gamma}(Y) - H_f(Y | X) \}.
\]

Since the above inequality holds for all $0 < \gamma < 1 - \varepsilon$, we have:

\[
C_e \leq \inf_{0 \leq \gamma < 1 - \varepsilon} \{ \sup_x \{ H_{e+\gamma}(Y) - H_f(Y | X) \} \}.
\]

By the symmetry of the channel, $H_f(Y | X) = H_f(Z)$ which is independent of $X$. Hence,

\[
C_e \leq \inf_{0 \leq \gamma < 1 - \varepsilon} \{ \log 2 - H_f(Z) \} = \inf_{0 \leq \gamma < 1 - \varepsilon} \{ 1 - H_f(Z) \},
\]

where the last step follows by taking a Bernoulli uniform input. Since $1 - H_f(Z)$ is non-increasing in $\gamma$,

\[
C_e \leq 1 - H_{1 - \varepsilon}(Z).
\]

(Note that the superscript "-" indicates a strict inequality in the definition of $H_f(\cdot)$; this is consistent with the condition $\gamma < 1$).

On the other hand, we can derive the lower bound to $C_e$ by choosing a Bernoulli uniform input in (15). We thus obtain

\[
1 - H_{1 - \varepsilon}(Z) \leq C_e \leq 1 - H_{1 - \varepsilon}(Z).
\]

Note that there are actually two upper bounds (16). In this example, the first upper bound $1 - H_{1 - \varepsilon}(Z)$ (which is no less than $1 - H_{1 - \varepsilon}(Z)$) is a looser upper bound, and hence, can be omitted. In addition, we demonstrate in the above derivation that the computation of the upper bound to $C_e$ involves in general the infimum operation over the parameter $\gamma$. Therefore, if the optimizing input distribution does not have a "nice" property (such as independence and uniformity), then the computation of (17) may be complicated in general.

**Remark:**

An alternative method to compute $C_e$ is to derive the channel sup-spectrum in terms of the inf-spectrum of the noise process. Under the optimizing equally likely Bernoulli input $X^*$ we can write

\[
\overline{\ell}(x^*;y)(\theta) = \lim_{n \to \infty} \sup \{ \frac{1}{n} \log P_{Y^t}(Y^t | X^n) \leq \theta \}
\]

\[
= \lim_{n \to \infty} \sup \{ \frac{1}{n} \log P_{Y^t}(Z^n) - \frac{1}{n} \log P_{Y^t}(Y^n) \leq \theta \}
\]

\[
= \lim_{n \to \infty} \sup \{ \frac{1}{n} \log P_{Y^t}(Z^n) \leq \theta - 1 \}
\]

\[
= \lim_{n \to \infty} \sup \{ \frac{1}{n} \log P_{Y^t}(Z^n) \geq 1 - \theta \}
\]

\[
= 1 - h_Z(1 - \theta).
\]

Hence,

\[
I_e(X^*;Y) = \sup \{ \theta : 1 - h_Z(1 - \theta) \leq \varepsilon \} = \sup \{ \theta : h_Z(1 - \gamma) \geq 1 - \varepsilon \}
\]

\[
= \sup \{ (1 - \beta) : h_Z(\beta^t) \geq 1 - \varepsilon \}
\]

\[
= 1 + \sup \{ (1 - \beta) : h_Z(\beta^t) \geq 1 - \varepsilon \}
\]

\[
= 1 - \inf \{ \beta : h_Z(\beta^t) \geq 1 - \varepsilon \}
\]

\[
= 1 - \sup \{ \beta : h_Z(\beta^t) < 1 - \varepsilon \}
\]

\[
= 1 - H(1 - \varepsilon)(Z).
\]

Similarly,

\[
I_e(X^*;Y) = 1 - H(1 - \varepsilon)(Z).
\]

Therefore,

\[
1 - H(1 - \varepsilon)(Z) = I_e(X^*;Y) \leq C_e \leq I_e(X^*;Y) = 1 - H(1 - \varepsilon)(Z).
\]

**Example 1.**

Let $Z$ be an all-zero sequence with probability $\beta$ and Bernoulli (with parameter $p$) with probability $1 - \beta$. Then the sequence of random variables $(1/n)h_{2^n}(Z^n)$ converges to atoms $0$ and $h_{2}(p) = -p \log p - (1 - p) \log (1 - p)$ with respective masses $\beta$ and $1 - \beta$. The resulting $h_Z(\theta)$ is depicted in Fig. 2. From (18), we obtain \( \overline{\ell}(x^*;y)(\theta) \) as shown in Fig. 3.

Therefore,

\[
C_e = \begin{cases} 
1 - h_2(\beta), & \text{if } 0 < \varepsilon < 1 - \beta; \\
1, & \text{if } 1 - \beta < \varepsilon < 1.
\end{cases}
\]

When $\varepsilon = 1 - \beta$, $C_e$ lies somewhere between $1 - h_2(p)$ and 1.

**Example 2.**

If $Z$ is a non-stationary binary independent
sequence with $Pr\{Z=1\}=p_i$, then by the uniform boundedness (in $i$) of the variance of random variable $-\log P_{Z_i}(Z_i)$, namely,

$$\text{Var}\{-\log P_{Z_i}(Z_i)\} < E\{(\log P_{Z_i}(Z_i))^2\}$$

$$\leq \sup_{0<p_i<1} p_i(\log p_i)^2+(1-p_i)(\log(1-p_i))^2$$

$$\leq 1,$$

we have (by Chebyshev’s inequality)

$$\Pr\left|\left(-\frac{1}{n}\log P_{Z}(Z^n)-\frac{1}{n} \sum_{i=1}^{n} H(Z_i)\right)\right|<\gamma \rightarrow 0,$$

for any $\gamma>0$. Therefore, $H(1-\varepsilon)(Z)$ is independent of $\varepsilon$, and $C_\varepsilon$ is equal to 1 minus the largest cluster point of $(1/n)\sum_{i=1}^{n} H(Z_i)$, i.e.,

$$\overline{H}_{[1-\varepsilon]}(Z)=\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} H(Z_i),$$

and

$$C_\varepsilon=1-\overline{H}(Z)=1-\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} H(Z_i),$$

where $H(Z_i)=h_b(p_i)$. This result is illustrated in Figs. 4 and 5.

**V. CONCLUSIONS**

In light of the work of Han and Verdú in [10] and [11], generalized entropy, mutual-information, and divergence rates are proposed. The properties of each of these information quantities are analyzed, and examples illustrating the computation of the $\varepsilon$-capacity of channels with arbitrary additive noise are presented.

In [3], we use these information measures to prove a generalized version of the Asymptotic Equipartition Property (AEP) and general source coding and hypothesis testing theorems.

**ACKNOWLEDGMENT**

The authors would like to thank Prof. S. Verdú for his valuable advice and constructive criticism which helped improve the paper.

**REFERENCES**


9. Gray, R.M. *Entropy and Information Theory*.


Discussions of this paper may appear in the discussion section of a future issue. All discussions should be submitted to the Editor-in-Chief.

*Manuscript Received: June 25, 1997
Revision Received: Jan. 11,1998
and Accepted: Jan. 24, 1998*