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

Po-Ning Chen [a] & Fady Alajaji [b]

[a] Dept. of Communications Engineering , National Chiao Tung University , Hsin Chu, Taiwan 300, R.O.C.

[b] Dept. of Mathematics and Statistics , Queen's University , Kingston, Ontario, K7L 3N6, Canada
Published online: 03 Mar 2011.

PLEASE SCROLL DOWN FOR ARTICLE

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

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 ε-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 [7]

$$H(X) \overset{\Delta}{=} -\sum_x P_X(x)\log_2 P_X(x) = E_{P_X}[-\log P_X(x)].$$

$H(X)$ is a measure of the average amount of uncertainty in X. The divergence, on the other hand, measures the relative distance between the distributions of two random variables X and $\hat{X}$ that are defined on the same alphabet :

$$D(X \| \hat{X}) \overset{\Delta}{=} E_{P_X}[\log_2 \frac{P_X(x)}{P_{\hat{X}}(x)}].$$

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 $P_X P_Y$:

$$I(X;Y) \overset{\Delta}{=} D(P_{XY} \| P_X P_Y) = E_{P_{XY}}[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}].$$

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

---

*Correspondence addressee

$$H(X) \overset{\Delta}{=} \lim_{n \to \infty} \frac{1}{n} E[-\log P_{X^n}(X^n)],$$

assuming the limit exists. Similarly the expressions for the divergence and mutual information rates are given by

$$D(X \| \hat{X}) \overset{\Delta}{=} \lim_{n \to \infty} \frac{1}{n} E[\log \frac{P_{X^n}(X^n)}{P_{\hat{X}^n}(X^n)}],$$

and

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

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 *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 *sup-entropy rate* $\overline{H}(X)$, 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 ε-capacity C_ε and capacity[1] C of arbitrary single-user channels without feedback (not necessarily information stable, stationary, ergodic, etc.). More specifically, they show that

$$\sup_X \sup \{R: F_X(R) < \varepsilon\} \le C_\varepsilon \le \sup_X \sup \{R: F_X(R) \le \varepsilon\},$$

and

$$C = \sup_X \underline{I}(X;Y),$$

where

$$F_X(R) \overset{\Delta}{=} \lim_{n \to \infty} \sup \Pr[(1/n) i_{X^n Y^n}(X^n;Y^n) \le R],$$

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

$$i_{X^n Y^n}(X^n; Y^n) = \log \frac{P_{Y^n | X^n}(y^n | x^n)}{P_{Y^n}(y^n)},$$

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

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

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

Similarly, its *limsup in probability* is the smallest extended real number $\overline{U}$ such that for all $\xi > 0$,

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

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

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

and

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

---

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

In other words, C_ε 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 ε, for sufficiently large blocklengths. Furthermore, C represents the largest rate at which information can be transmitted over the channel with asymptotically *vanishing* error probability.

We can observe however that there might exist cases of interest where *only* the liminfs of the probabilities in (3) and (4) are equal to zero; while the limsups do *not* vanish. There are also other cases where *both* the liminfs and limsups in (3)-(4) do not vanish; but they are upper bounded by a prescribed threshold. Furthermore, there are situations where the interval $[\underline{U}, \overline{U}]$ 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 $[\underline{U}, \overline{U}]$ 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 ε-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 ε-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

$$\underline{u}(\theta) \stackrel{\Delta}{=} \liminf_{n \to \infty} \Pr[A_n \le \theta],$$

and

$$\overline{u}(\theta) \stackrel{\Delta}{=} \limsup_{n \to \infty} \Pr[A_n \le \theta].$$

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 $A_n - \Pr\{A_n \le \theta\}$—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 $0 \le \delta \le 1$, the quantiles $\underline{U}_\delta$ and $\overline{U}_\delta$ of the sup-spectrum and the inf-spectrum are defined by[3]

$$\underline{U}_\delta \stackrel{\Delta}{=} \begin{cases} -\infty & \text{if } \{\theta : \overline{u}(\theta) \le \delta\} = \varnothing, \\ \sup\{\theta : \overline{u}(\theta) \le \delta\}, & \text{otherwise}, \end{cases}$$

and

$$\overline{U}_\delta \stackrel{\Delta}{=} \begin{cases} -\infty & \text{if } \{\theta : \underline{u}(\theta) \le \delta\} = \varnothing, \\ \sup\{\theta : \underline{u}(\theta) \le \delta\}, & \text{otherwise}, \end{cases}$$

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

Note that the liminf in probability $\underline{U}$ and the limsup in probability $\overline{U}$ of A_n satisfy

$$\underline{U} = \underline{U}_0$$

and

$$\overline{U} = \overline{U}_{1^-},$$

respectively, where the superscript "–" denotes a strict inequality in the definition of $\overline{U}_{1^-}$; i.e.,

$$\overline{U}_\delta \stackrel{\Delta}{=} \sup\{\theta : \underline{u}(\theta) < \delta\}.$$

Note also that

$$\underline{U} \le \underline{U}_\delta \le \overline{U}_\delta \le \overline{U}.$$

Remark that $\underline{U}_\delta$ and $\overline{U}_\delta$ always exist. Furthermore, if $\underline{U}_\delta = \overline{U}_\delta \ \forall \delta \in [0,1]$, 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 δ-inf-entropy-rate $\underline{H}_\delta(X)$*

---

[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 spectrums $\underline{u}(\cdot)$ 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 $\underline{u}(\cdot)$ 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 δ-*sup-entropy-rate* $\overline{H}_\delta(X)$ as described in Table 1. Note that the *inf-entropy-rate* $\underline{H}(X)$ and the *sup-entropy-rate* $\overline{H}(X)$ introduced in [10] are special cases of the δ-inf/sup-entropy rate measures:

$$\underline{H}(X)=\underline{H}_0(X), \text{ and } \overline{H}(X)= \overline{H}_1\text{-}(X).$$

Analogously, for an arbitrary channel $W\overset{\Delta}{=}P_{Y|X}$ with input X and output Y (or respectively for two observations X and $\hat{X}$), if we replace A_n by the normalized information density (resp. by the normalized log-likelihood ratio), we get the δ-inf/sup-information rates (resp. δ-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 $\overline{u}$ $(\cdot)$ and $\underline{u}(\cdot)$ denote respectively the sup-spectrum and inf-spectrum of $\{A_n\}_{n=1}^\infty$. Similarly, let $\overline{v}$ $(\cdot)$ and $\underline{v}(\cdot)$ denote respectively the sup-spectrum and inf-spectrum of $\{B_n\}_{n=1}^\infty$. Define $\underline{U}_\delta \overset{\Delta}{=} \sup\{\theta: \overline{u}(\theta)\le\delta\}$, $\overline{U}_\delta \overset{\Delta}{=} \sup\{\theta:\underline{u}(\theta)\le\delta\}$, $\underline{V}_\delta=\sup\{\theta: \overline{v}(\theta)\le\delta\}$, $\overline{V}_\delta \overset{\Delta}{=} \sup\{\theta:\underline{v}(\theta)\le\delta\}$,

$$(\underline{U+V})_{\delta+\gamma} \overset{\Delta}{=} \sup\{\theta: \overline{(u+v)}(\theta)\le\delta+\gamma\},$$

$$(\overline{U+V})_{\delta+\gamma} \overset{\Delta}{=} \sup\{\theta: \underline{(u+v)}(\theta)\le\delta+\gamma\},$$

$$\overline{(u+v)}(\theta) \overset{\Delta}{=} \limsup_{n\to\infty} \Pr\{A_n+B_n\le\theta\},$$

and

$$\underline{(u+v)}(\theta) \overset{\Delta}{=} \liminf_{n\to\infty} \Pr\{A_n+B_n\le\theta\},$$

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 $\delta\ge0$, $\gamma\ge0$, and $1\ge\delta+\gamma$,

$$(\underline{U+V})_{\delta+\gamma}\ge\underline{U}_\delta+\underline{V}_\gamma, \tag{5}$$

and

$$(\overline{U+V})_{\delta+\gamma}\ge \underline{U}_\delta+ \overline{V}_\gamma. \tag{6}$$

3. For $\delta\ge0$, $\gamma\ge0$, and $1>\delta+\gamma$,

$$(\underline{U+V})_\delta\le\underline{U}_{\delta+\gamma}+ \overline{V}_{(1-\gamma)^-}, \tag{7}$$
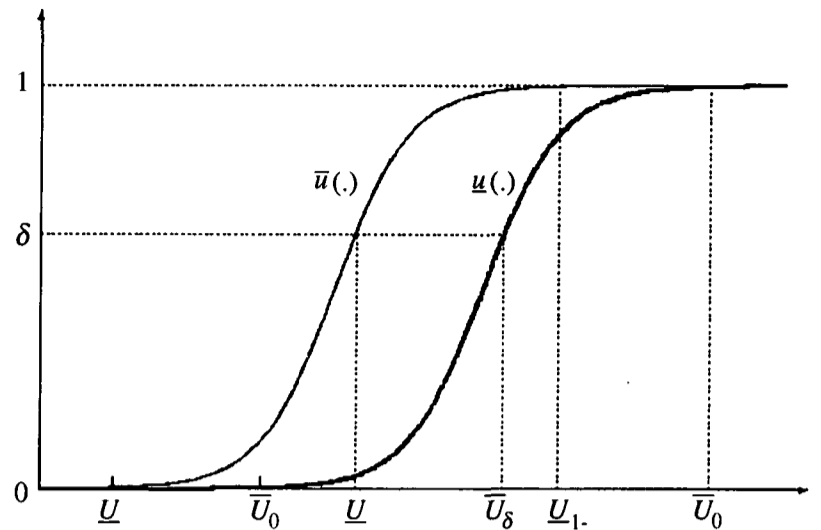
and

Fig. 1. The asymptotic CDFs of a sequence of random variables $\{A_n\}_{n=1}^\infty$. $\overline{u}$ $(\cdot)$=sup-spectrum of A_n; $\underline{u}(\cdot)$=inf-spectrum of A_n.

$$(\overline{U+V})_\delta\le \overline{U}_{\delta+\gamma}+\overline{V}_{(1-\gamma)^-}. \tag{8}$$

Proof:

The proof of property 1 follows directly from the definitions of $\underline{U}_\delta$ and $\overline{U}_\delta$ and the fact that the inf-spectrum and the sup-spectrum are non-decreasing in δ.

To show (5), we first observe that

$$\Pr\{A_n+B_n\le\underline{U}_\delta+\underline{V}_\gamma\}\le\Pr\{A_n\le\underline{U}_\delta\}+\Pr\{B_n\le\underline{V}_\gamma\}.$$

Then

$$\limsup_{n\to\infty}\Pr\{A_n+B_n\le\underline{U}_\delta+\underline{V}_\gamma\}$$

$$\le\limsup_{n\to\infty}(\Pr\{A_n\le\underline{U}_\delta\}+\Pr\{B_n\le\underline{V}_\gamma\})$$

$$\le\limsup_{n\to\infty}\Pr\{A_n\le\underline{U}_\delta\}+\limsup_{n\to\infty}\Pr\{B_n\le\underline{V}_\gamma\}$$

$$\le\delta+\gamma,$$

which, by definition of $(\underline{U+V})_{\delta+\gamma}$, yields (5). Similarly, we have

$$\Pr\{A_n+B_n\le\underline{U}_\delta+\overline{V}_\gamma\}\le\Pr\{A_n\le\underline{U}_\delta\}+\Pr\{B_n\le\overline{V}_\gamma\}.$$

Then

$$\liminf_{n\to\infty}\Pr\{A_n+B_n\le\underline{U}_\delta+\overline{V}_\gamma\}$$

$$\le\liminf_{n\to\infty}(\Pr\{A_n\le\underline{U}_\delta\}+\Pr\{B_n\le\overline{V}_\gamma\})$$

$$\le\limsup_{n\to\infty}\Pr\{A_n\le\underline{U}_\delta\}+\liminf_{n\to\infty}\Pr\{B_n\le\overline{V}_\gamma\}$$

$$\le\delta+\gamma,$$

which, by definition of $(\overline{U+V})_{\delta+\gamma}$, proves (6).

To show (7), we remark from (5) that $(\underline{U+V})_\delta+$
$(\underline{-V})_\gamma\le(\underline{U+V-V})_{\delta+\gamma}=\underline{U}_{\delta+\gamma}$ Hence,

Table 1. Generalized information measures where $\delta \in [0,1]$.

Entropy Measures	
System	Arbitrary Source X
A_n: Norm. Entropy Density	$(1/n)\,h_{X^n}(X^n) \overset{\Delta}{=} -(1/n)\log P_{X^n}(X^n)$
Entropy Sup-Spectrum	$\overline{h}_X(\theta) \overset{\Delta}{=} \limsup_{n\to\infty} \Pr\{(1/n)\,h_{X^n}(X^n) \le \theta\}$
Entropy Inf-Spectrum	$\underline{h}_X(\theta) \overset{\Delta}{=} \liminf_{n\to\infty} \Pr\{(1/n)\,h_{X^n}(X^n) \le \theta\}$
δ-Inf-Entropy Rate	$\underline{H}_\delta(X) \overset{\Delta}{=} \sup\{\theta\colon \overline{h}_X(\theta) \le \delta\}$
δ-Sup-Entropy Rate	$\overline{H}_\delta(X) \overset{\Delta}{=} \sup\{\theta\colon \underline{h}_X(\theta) \le \delta\}$
Sup-Entropy Rate	$\overline{H}(X) \overset{\Delta}{=} \overline{H}_1{}^-(X)$
Inf-Entropy Rate	$\underline{H}(X) \overset{\Delta}{=} \underline{H}_0(X)$

Mutual Information Measures	
System	Arbitrary channel $W \overset{\Delta}{=} P_{Y\mid X}$ with input X and output Y
A_n: Norm. Information Density	$(1/n)\,i_{(X^n,Y^n)}(X^n;Y^n) \overset{\Delta}{=} (1/n)\log \dfrac{dP_{X^nY^n}}{d(P_{X^n}\times P_{Y^n})}(X^n,Y^n)$
Information Sup-Spectrum	$\overline{i}_{(X,Y)}(\theta) \overset{\Delta}{=} \limsup_{n\to\infty} \Pr\{(1/n)\,i_{(X^n,Y^n)}(X^n;Y^n) \le \theta\}$
Information Inf-Spectrum	$\underline{i}_{(X,Y)}(\theta) \overset{\Delta}{=} \liminf_{n\to\infty} \Pr\{(1/n)\,i_{(X^n,Y^n)}(X^n;Y^n) \le \theta\}$
δ-Inf-Information Rate	$\underline{I}_\delta(X;Y) \overset{\Delta}{=} \sup\{\theta\colon \overline{i}_{(X,Y)}(\theta) \le \delta\}$
δ-Sup-Information Rate	$\overline{I}_\delta(X;Y) \overset{\Delta}{=} \sup\{\theta\colon \underline{i}_{(X,Y)}(\theta) \le \delta\}$
Sup-Information Rate	$\overline{I}(X;Y) \overset{\Delta}{=} \overline{I}_1{}^-(X;Y)$
Inf-Information Rate	$\underline{I}(X;Y) \overset{\Delta}{=} \underline{I}_0(X;Y)$

Divergence Measures	
System	Arbitrary sources X and $\hat{X}$
A_n: Norm Log-Likelihood Ratio	$(1/n)\,d_{X^n}(X^n \parallel \hat{X}^n) \overset{\Delta}{=} (1/n)\log[dP_{X^n}/dP_{\hat{X}^n}](X^n)$
Divergence Sup-Spectrum	$\overline{d}_{X\mid\hat{X}}(\theta) \overset{\Delta}{=} \limsup_{n\to\infty} \Pr\{(1/n)\,d_{X^n}(X^n \parallel \hat{X}^n) \le \theta\}$
Divergence Inf-Spectrum	$\underline{d}_{X\mid\hat{X}}(\theta) \overset{\Delta}{=} \liminf_{n\to\infty} \Pr\{(1/n)\,d_{X^n}(X^n \parallel \hat{X}^n) \le \theta\}$
δ-Inf-Divergence Rate	$\underline{D}_\delta(X\parallel\hat{X}) \overset{\Delta}{=} \sup\{\theta\colon \overline{d}_{X\parallel\hat{X}}(\theta) \le \delta\}$
δ-Sup-Divergence Rate	$\overline{D}_\delta(X\parallel\hat{X}) \overset{\Delta}{=} \sup\{\theta\colon \underline{d}_{X\parallel\hat{X}}(\theta) \le \delta\}$
Sup-Divergence Rate	$\overline{D}(X\parallel\hat{X}) \overset{\Delta}{=} \overline{D}_1{}^-(X\parallel\hat{X})$
Inf-Divergence Rate	$\underline{D}(X\parallel\hat{X}) \overset{\Delta}{=} \underline{D}_0(X\parallel\hat{X})$

$(\underline{U+V})_\delta \le \underline{U}_{\delta+\gamma}{}^-(\underline{-V})_\gamma$

(Note that the cases $\delta+\gamma=1$ or $\gamma=1$ are not allowed here because they result in $\underline{U}_1 = \underline{-V}_1 = \infty$, and the subtraction of two infinite terms is undefined. That is why the condition for property 2, $1 \le \delta+\gamma$, is replaced by $1 > \delta+\gamma$ in property 3.)

The proof is completed by showing that

$$-(\underline{-V})_\gamma \le \overline{V}_{(1-\gamma)}{}^- . \tag{9}$$

By definition,

$$\overline{(-v)}(\theta) \overset{\Delta}{=} \limsup_{n\to\infty} \Pr\{-B_n \le \theta\}$$

$$= 1 - \lim_{n \to \infty} \inf \Pr\{B_n < -\theta\}$$

$$= 1 - \underline{v}(-\theta^+).$$

So $\underline{v}(-\theta^+) = 1 - (\overline{-v})(\theta)$. Then

$$\overline{V}_{(1-\gamma)^-} \overset{\Delta}{=} \sup\{\theta : \underline{v}(\theta) < 1 - \gamma\}$$

$$\geq \sup\{\theta : \underline{v}(\theta^-) < 1 - \gamma\}$$

$$= \sup\{-\theta : \underline{v}(-\theta^+) < 1 - \gamma\}$$

$$= \sup\{-\theta : 1 - (\overline{-v})(\theta) < 1 - \gamma\}$$

$$= -\inf\{\theta : (\overline{-v})(\theta) > \gamma\}$$

$$= -\sup\{\theta : (\overline{-v})(\theta) \leq \gamma\}$$

$$= -(\overline{-V})_\gamma$$

where the inequality follows from $\underline{v}(\theta) \geq \underline{v}(\theta^-)$. Finally, to show (8), we observe from (6) that $(\overline{U+V})_\delta + (\underline{-V})_\gamma \leq \overline{(U+V-V)}_{\delta+\gamma} = \overline{U}_{\delta+\gamma}$. Hence,

$$(\overline{U+V})_\delta \leq \overline{U}_{\delta+\gamma} - (\underline{-V})_\gamma.$$

Using (9), we have the desired result. ∎

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

$$(\underline{U+V}) \geq \underline{U} + \underline{V}, \text{ and } (\overline{U+V}) \leq \underline{U} + \overline{V},$$

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

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

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 δ, γ, $\delta + \gamma \in [0,1)$, the following statements hold.

1. $\overline{H}_\delta(X) \geq 0$. $\overline{H}_\delta(X) = 0$ if and only if the sequence $\{X^n = (X_1^{(n)}, ..., X_n^{(n)})\}_{n=1}^\infty$ is ultimately deterministic (in probability).

   (This property also applies to $\underline{H}_\delta(X)$, $\overline{I}_\delta(X;Y)$, $\underline{I}_\delta(X;Y)$, $\overline{D}_\delta(X\|Y)$, and $\underline{D}_\delta(Y\|\hat{X})$.)

2. $\underline{I}_\delta(X;Y) = \underline{I}_\delta(Y;X)$ and $\overline{I}_\delta(X;Y) = \overline{I}_\delta(Y;X)$.

3.

$$\underline{I}_\delta(X;Y) \leq \underline{H}_{\delta+\gamma}(Y) - \underline{H}_\gamma(Y \mid X), \tag{10}$$

$$\underline{I}_\delta(X;Y) \leq \overline{H}_{\delta+\gamma}(Y) - \overline{H}_\gamma(Y \mid X), \tag{11}$$

$$\overline{I}_\gamma(X;Y) \leq \overline{H}_{\delta+\gamma}(Y) - \underline{H}_\delta(Y \mid X), \tag{12}$$

$$\underline{I}_{\delta+\gamma}(X;Y) \geq \underline{H}_\delta(Y) - \overline{H}_{(1-\gamma)^-}(Y \mid X), \tag{13}$$

and

$$\overline{I}_{\delta+\gamma}(X;Y) \geq \overline{H}_\delta(Y) - \overline{H}_{(1-\gamma)^-}(Y \mid X) \tag{14}$$

4. $0 \leq \underline{H}_\delta(X) \leq \overline{H}_\delta(X) \leq \log|X|$, where each $X_i^{(n)} \in X$, $i=1, ..., n$ and $n=1,2,...,$ and X is finite.

5. $\underline{I}_\delta(X,Y;Z) \geq \underline{I}_\delta(X;Z)$.

Proof:

Property 1 holds because

$$\Pr\{-\frac{1}{n} \log P_{X^n}(X^n) < 0\} = 0,$$

$$\Pr\{\frac{1}{n} \log \frac{dP_{X^n}}{dP_{\hat{X}^n}}(X^n) < -\delta\} \leq \exp\{-\delta n\}.$$

and

$$\Pr\{\frac{1}{n} \log \frac{dP_{X^n Y^n}}{d(P_{X^n} \times P_{Y^n})}(X^n, Y^n) < -\delta\} \leq \exp\{-\delta n\}.$$

Property 2 is an immediate consequence of the definition.

To show the inequalities in property 3 we first remark that

$$\frac{1}{n} h_{Y^n}(Y^n) = \frac{1}{n} i_{(X^n,Y^n)}(X^n,Y^n) + \frac{1}{n} h_{(X^n,Y^n)}(Y^n \mid X^n),$$

where $(1/n)h_{(X^n, Y^n)}(Y^n \mid X^n) \overset{\Delta}{=} -(1/n)\log P_{Y^n \mid X^n}(Y^n \mid X^n)$. 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 $\overline{H}_\delta(\cdot)$ is non-decreasing in δ: $\overline{H}_\delta(X) \leq \overline{H}_{1^-}(X) = \overline{H}(X)$, and that $\overline{H}(X)$ 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

$$\frac{1}{n} i_{(X^n,Y^n,Z^n)}(X^n,Y^n,Z^n) = \frac{1}{n} i_{(X^n,Z^n)}(X^n,Z^n)$$

$$+ \frac{1}{n} i_{(X^n,Y^n,Z^n)}(Y^n;Z^n \mid X^n).$$

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_1^n and X_3^n are conditionally independent given X_2^n. Then

$$\underline{I}_\delta(X_1;X_3) \leq \underline{I}_\delta(X_1;X_2).$$

Proof:

By property 5, we get

$$\underline{I}_\delta(X_1;X_3) \leq \underline{I}_\delta(X_1;X_2,X_3) = \underline{I}_\delta(X_1;X_2),$$

where the equality holds because

$$\frac{1}{n}\log\frac{dP_{X_1^n X_2^n X_3^n}}{d(P_{X_1^n} \times P_{X_2^n X_3^n})}(X_1^n,X_2^n,X_3^n) = \frac{1}{n}\log\frac{dP_{X_1^n X_2^n}}{d(P_{X_1^n} \times P_{X_2^n})}(X_1^n,X_2^n).$$

∎

Lemma 4. (Optimality of independent inputs)

Fix $\delta \in [0,1)$. Consider a finite alphabet, discrete memoryless channel – i.e., $P_{Y^n|X^n} = \Pi_{i=1}^n P_{Y_i|X_i}$, for all n. For any input X and its corresponding output Y,

$$\underline{I}_\delta(X,Y) \leq \underline{I}_\delta(\overline{X};\overline{Y}) = \underline{I}(\overline{X};\overline{Y}),$$

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

Proof:

First, we observe that

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

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

In other words,

$$\frac{1}{n}\log\frac{dP_{X^n Y^n}}{d(P_{X^n} \times P_{Y^n})}(X^n,Y^n) + \frac{1}{n}\log\frac{dP_{Y^n}}{dP_{\overline{Y}^n}}(X^n,Y^n)$$

$$= \frac{1}{n}\log\frac{dP_{\overline{X}^n \overline{Y}^n}}{d(P_{\overline{X}^n} \times P_{\overline{Y}^n})}(X^n,Y^n).$$

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

$$\overline{z}(\theta) \triangleq \limsup_{n \to \infty} P_{X^n Y^n}\{\frac{1}{n}\log\frac{dP_{\overline{X}^n \overline{Y}^n}}{d(P_{\overline{X}^n} \times P_{\overline{Y}^n})}(X^n,Y^n) \leq \theta\},$$

and

$$\underline{Z}_\delta(\overline{X},\overline{Y}) \triangleq \sup\{\theta: \overline{z}(\theta) \leq \delta\},$$

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

$$\underline{Z}_\delta(\overline{X},\overline{Y}) \geq \underline{I}_\delta(X,Y) + \underline{D}(Y\|\overline{Y}) \geq \underline{I}_\delta(X;Y),$$

since $\underline{D}(Y\|\overline{Y}) \geq 0$ by property 1 of Lemma 2.

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

$$\underline{I}(\overline{X};\overline{Y}) = \underline{I}_\delta(\overline{X};\overline{Y}) \text{ and } \underline{Z}(\overline{X};\overline{Y}) = \underline{Z}_\delta(\overline{X};\overline{Y}).$$

It finally remains to show that

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

which is proved in [11, Theorem 10].

∎

IV. EXAMPLES FOR THE COMPUTATION OF ε-CAPACITY

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

$$\sup_X \underline{I}_{\varepsilon^-}(\overline{X};\overline{Y}) \leq C_\varepsilon \leq \sup_X \underline{I}_\varepsilon(\overline{X};\overline{Y}),$$

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

We now provide examples for the computation of C_ε. 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 $\mathcal{X}=\mathcal{Y}=\{0,1\}$, 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 ε-capacity we use the results of property 3 in Lemma 2:

$$\underline{I}_{\varepsilon^-}(\overline{X};\overline{Y}) \geq \underline{H}_0(Y) - \overline{H}_{(1-\varepsilon^-)}(Y|X) = \underline{H}_0(Y) - \overline{H}_{(1-\varepsilon)}(Y|X),$$

(15)

and

$$\underline{I}_\varepsilon(\underline{X};\underline{Y}) \leq \min\{\underline{H}_{\varepsilon^+}(Y) - \underline{H}_\gamma(Y|X), \overline{H}_{\varepsilon^+}(Y) - \overline{H}_\gamma(Y|X)\},$$

(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.

$$C_\varepsilon \leq \sup_X \underline{I}_\varepsilon(X;Y)$$

$$\leq \sup_X \{\overline{H}_{\varepsilon+\gamma}(Y) - \overline{H}_\gamma(Y \mid X)\}.$$

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

$$C_\varepsilon \leq \inf_{0 \leq \gamma < 1-\varepsilon} \sup_X \{\overline{H}_{\varepsilon+\gamma}(Y) - \overline{H}_\gamma(Y \mid X)\}$$

$$\leq \inf_{0 \leq \gamma < 1-\varepsilon} \{\sup_X \overline{H}_{\varepsilon+\gamma}(Y) - \inf_X \overline{H}_\gamma(Y \mid X)\}.$$

By the symmetry of the channel, $\overline{H}_\gamma(Y|X) = \overline{H}_\gamma(Z)$ which is independent of X. Hence,

$$C_\varepsilon \leq \inf_{0 \leq \gamma < 1-\varepsilon} \{\sup_X \overline{H}_{\varepsilon+\gamma}(Y) - \overline{H}_\gamma(Z)\}$$

$$\leq \inf_{0 \leq \gamma < 1-\varepsilon} \{\log_2 2 - \overline{H}_\gamma(Z)\} = \inf_{0 \leq \gamma < 1-\varepsilon} \{1 - \overline{H}_\gamma(Z)\},$$

where the last step follows by taking a Bernoulli uniform input. Since $1 - \overline{H}_\gamma(Z)$ is non-increasing in γ,

$$C_\varepsilon \leq 1 - \overline{H}_{(1-\varepsilon)^-}(Z).$$

(Note that the superscript "−" indicates a strict inequality in the definition of $\overline{H}_\gamma(\cdot)$; this is consistent with the condition $\gamma + \varepsilon < 1$.)

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

$$1 - \overline{H}_{(1-\varepsilon)}(Z) \leq C_\varepsilon \leq 1 - \overline{H}_{(1-\varepsilon)^-}(Z).$$

Note that there are actually two upper bounds (16). In this example, the first upper bound $1 - \underline{H}_{(1-\varepsilon)^-}(Z)$ (which is no less than $1 - \overline{H}_{(1-\varepsilon)^-}(Z)$) 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_ε involves in general the infimum operation over the parameter γ. 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_ε 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

$$\overline{i}_{(X^*;Y)}(\theta) = \limsup_{n \to \infty} \Pr \{\frac{1}{n}\log \frac{P_{Y^n|X^n}(Y^n \mid X^n)}{P_{Y^n}(Y^n)} \leq \theta\}$$

$$= \limsup_{n \to \infty} \Pr \{\frac{1}{n}\log P_{Z^n}(Z^n) - \frac{1}{n}\log P_{Y^n}(Y^n) \leq \theta\}$$

$$= \limsup_{n \to \infty} \Pr \{\frac{1}{n}\log P_{Z^n}(Z^n) \leq \theta - 1\}$$

$$= \limsup_{n \to \infty} \Pr \{-\frac{1}{n}\log P_{Z^n}(Z^n) \geq 1 - \theta\}$$

$$= 1 - \underline{h}_Z((1 - \theta)^-).$$

Hence,

$$\underline{I}_\varepsilon(X^*;Y) = \sup\{\theta: 1 - \underline{h}_Z((1 - \theta)^-) \leq \varepsilon\}$$

$$= \sup\{\theta: \underline{h}_Z((1 - \theta)^-) \geq 1 - \varepsilon\}$$

$$= \sup\{(1 - \beta): \underline{h}_Z(\beta^-) \geq 1 - \varepsilon\}$$

$$= 1 + \sup\{(- \beta): \underline{h}_Z(\beta^-) \geq 1 - \varepsilon\}$$

$$= 1 - \inf\{\beta: \underline{h}_Z(\beta^-) \geq 1 - \varepsilon\}$$

$$= 1 - \sup\{\beta: \underline{h}_Z(\beta^-) < 1 - \varepsilon\}$$

$$= 1 - \overline{H}_{(1-\varepsilon)^-}(Z).$$

Similarly,

$$\overline{I}_\varepsilon(X^*;Y) = 1 - \overline{H}_{(1-\varepsilon)}(Z).$$

Therefore,

$$1 - \overline{H}_{(1-\varepsilon)}(Z) = \underline{I}_\varepsilon(X^*;Y) \leq C_\varepsilon \leq \underline{I}_\varepsilon(X^*;Y) = 1 - \overline{H}_{(1-\varepsilon)^-}(Z).$$

Example 1.

Let Z be an all-zero sequence with probability β and Bernoulli (with parameter p) with probability $1 - \beta$. Then the sequence of random variables $(1/n)h_{Z^n}(Z^n)$ converges to atoms 0 and $h_b(p) \overset{\Delta}{=} -p\log p - (1-p)\log(1-p)$ with respective masses β and $1 - \beta$. The resulting $\underline{h}_Z(\theta)$ is depicted in Fig. 2. From (18), we obtain $\overline{i}_{(X,Y)}(\theta)$ as shown in Fig. 3.

Therefore,

$$C_\varepsilon = \begin{cases} 1 - h_b(p), & \text{if } 0 < \varepsilon < 1 - \beta; \\ 1, & \text{if } 1 - \beta < \varepsilon < 1. \end{cases}$$

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

Example 2.

If Z is a non-stationary binary independent

Fig. 2. The spectrum of $(1/n)h_{Zn}(Z^n)$ for Example 1.

Fig. 3. The spectrum of $(1/n)i_{(X^n, Y^n)}(X^n; Y^n)$ for Example 1.

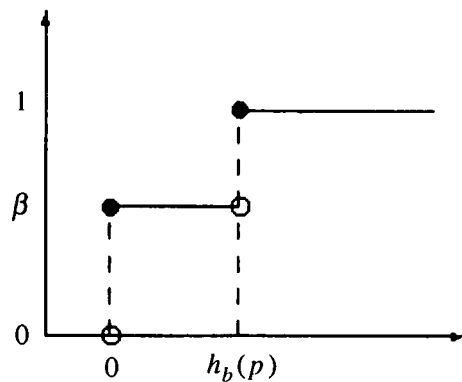Fig. 4. The spectrum of $(1/n)h_{Zn}(Z^n)$ for Example 2.
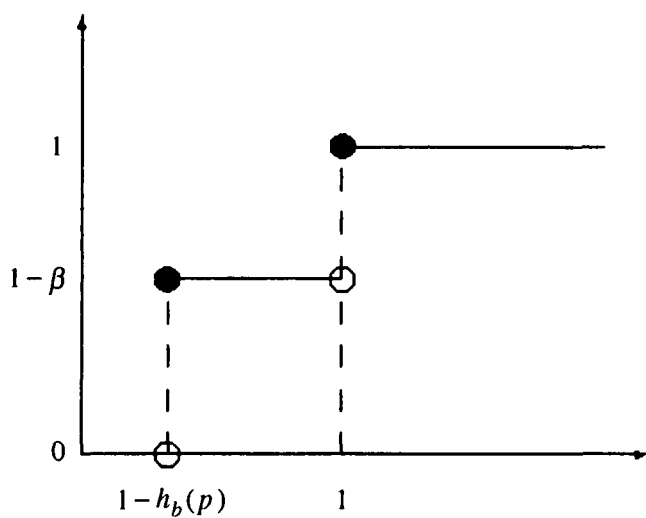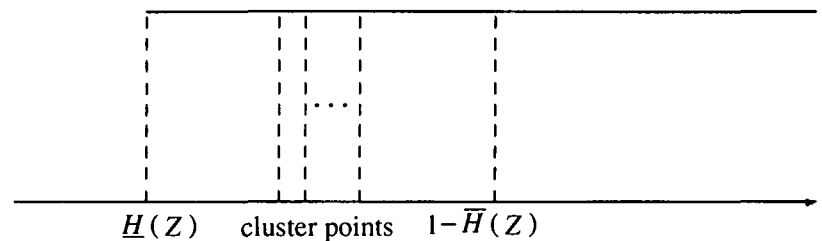

Fig. 5. The spectrum of $(1/n)i_{(X^n, Y^n)}(X^n; Y^n)$ for Example 2.

sequence with $\Pr\{Z_1=1\}=p_i$, then by the uniform boundedness (in i) of the variance of random variable $-\log P_{Z_i}(Z_i)$, namely,

$$\mathrm{Var}[-\log P_{Z_i}(Z_i)] \le E[(\log P_{Z_i}(Z_i))^2]$$

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

$$\le 1,$$

we have (by Chebyshev's inequality)

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

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

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

and

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

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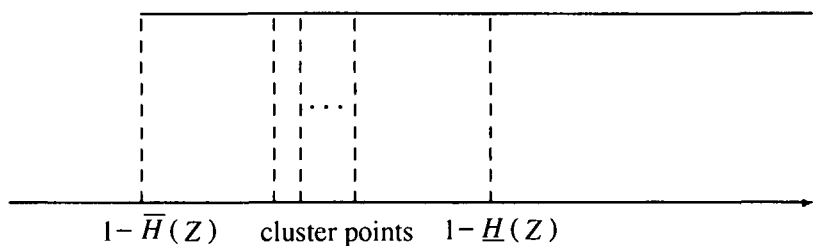
V. CONCLUSIONS

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 ε-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

REFERENCES

1. Billingsley, P. *Probability and Measure*, Wiley, New York (1986).
2. Blahut, R.E. *Principles and Practice of Information Theory*, Addison Wesley, Massachusetts (1988).
3. Chen P.-N. and F. Alajaji, "Generalized Source Coding Theorems and Hypothesis Testing: Part II – Operational limits," *Journal of the Chinese Institute of Engineers*, Vol. 21, No. 3, May (1998).
4. Chen P.-N. and F. Alajaji, "Strong Converse, Feedback Channel Capacity and Hypothesis Testing," *Journal of the Chinese Institute of Engineers*, Vol. 18, pp. 777-785, November 1995; also in Proceedings of CISS, John Hopkins Univ., MD, USA, March (1995).

5. Chen P.-N. and F. Alajaji, "The Reliability Function of Arbitrary Channels with and Without Feedback," *Proceedings of the 18'th Biennial Symposium on Communications*, Queen's University, Kingston, Ontario, June (1996).

6. Chen, P.-N. "General Formulas for the Neyman-Pearson type-II Error Exponent Subject to Fixed and Exponential type-I Error Bounds," *IEEE Transactions on Information Theory*, T-42(1): 316-323, January (1996).

7. Cover T.M. and J.A. Thomas, *Elements of Information Theory*, Wiley, New York (1991).

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

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

Springer-Verlag, New York (1990).

10. Han T.S. and S. Verdú, "Approximation Theory of Output Dtatistics," *IEEE Transactions on Information Theory*, IT-39(3): 752-772, May (1993).

11. Verdú S. and T.S. Han, "A General Formula for Channel Capacity," *IEEE Transactions on Information Theory*, IT-40(4) : 1147-1157, Jul. (1994).

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

來源編碼定理與檢定測試的一般定理：第一部份 －訊息量度

陳伯寧

國立交通大學電信工程系所

Fady Alajaji
Queen's University
Kingston, ON K7L 3N6, Canada

摘 要

本論文將介紹 ε-熵率、ε-相互訊息率與 ε-偏差率的表示式，這三個表示式是擴展Han與Verdú的極下／極上－熵／相互訊息／偏差率而得到的，式中包含了對應近似訊息機率譜。我們將分析它們的代數性質，舉例說明其於 ε-管道傳輸能力的計算上之應用。在此研究的第二部份，這三個公式將被用來證明區塊碼來源編碼一般定理與紐門皮爾森檢定測試第二類錯誤指數之一般公式。

關鍵詞：消息理論、近似平均分割性質、來源編碼定理、檢定測試、紐門皮爾森錯誤指數。