Generalization of Information Measures*

Po-Ning Chen† and Fady Alajaji‡

† Computer & Communication Research Laboratories,
Industrial Technology Research Institute, Taiwan 310, Republic of China
Email: poning@e0sun3.ccl.itri.org.tw

‡ Department of Mathematics & Statistics,
Queen's University, Kingston, ON K7L 3N6, Canada
Email: fady@polya.mast.queensu.ca

Abstract — **General formulas for entropy, mutual information, and divergence are established. It is revealed that these quantities are actually determined by three decisive sequences of random variables; which are, respectively, the normalized source information density, the normalized channel information density, and the normalized log-likelihood ratio. In terms of the ultimate cumulative distribution functions or spectrums of these random sequences, entropy, mutual information and divergence are respectively expressed in their most general form. In light of the newly defined quantities, general data compaction and data compression (source coding) theorems for block codes, and the Neyman-Pearson type-II error exponent subject to upper bounds on the type-I error probability are derived.**

I. Introduction

Entropy, divergence and mutual information are without a doubt the most important quantities in the fields of information and communication theory. Almost all information theoretical limits involve these quantities. The simplest expressions for entropy, divergence and mutual information are:

$H(X) \stackrel{\triangle}{=} E_{P_X}\left[-\log P_X(X)\right]$ for entropy,

$D(X\|\hat{X}) \stackrel{\triangle}{=} E_{P_X}\left[\log[dP_X/dP_{\hat{X}}](X)\right]$ for divergence,

and $I(X;Y) \stackrel{\triangle}{=} E_{P_{XY}}\left[\log[dP_{XY}/d(P_X \times P_Y)](X,Y)\right]$ for mutual information, where P_X, $P_{\hat{X}}$, P_{XY} and P_Y are distributions defined over some proper observation spaces, and the subscript of $E[\cdot]$ indicates the distribution employed in the expectation evaluation. These formulas have an operational significance only when the theoretical limits are considered under an independent and identically distributed environment. However, in more complicated cases such as when the statistics vary temporally, these formulas may no longer be valid and some kind of generalization is required.

A straightforward generalization of these quantities is to extend the simple per-letter formulas to their respective ultimate average rates:

$H(\boldsymbol{X}) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} E\left[-\log P_{X^n}(X^n)\right]$ for entropy rate,

$D(\boldsymbol{X}\|\hat{\boldsymbol{X}}) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} E\left[\log \frac{dP_{X^n}}{dP_{\hat{X}^n}}(X^n)\right]$ for divergence rate, and

$$I(\boldsymbol{X};\boldsymbol{Y}) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} E\left[\log \frac{dP_{X^nY^n}}{d(P_{X^n}\times P_{Y^n})}(X^n,Y^n)\right]$$

for mutual information rate, where $\boldsymbol{X}$ denotes a sequence of finite dimensional random variables; i.e.,

$$\boldsymbol{X} \stackrel{\triangle}{=} \left\{X^n = \left(X_1^{(n)},\ldots,X_n^{(n)}\right)\right\}_{n=1}^{\infty},$$

and similar definitions apply for $\hat{\boldsymbol{X}}$ and $\boldsymbol{Y}$. Therefore, Shannon's coding theorems can be generalized for systems where the above limits exist [1, 7].

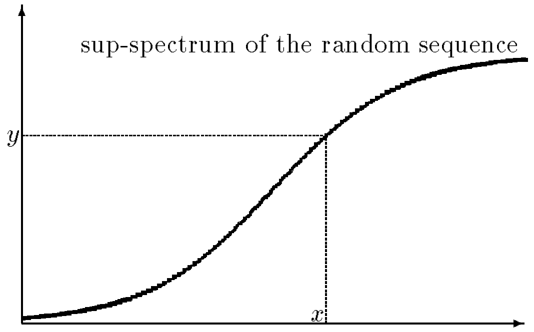
Constraints on the existence of the ultimate averages in the above formulas somewhat limit the usefulness of entropy, divergence and mutual information; in some specific situations, a lot of efforts is exerted to show their existence instead of examining the underlying theory itself. This leads us to raise the following question: If all assumptions such as memorylessness, stationarity, causality, ergodicity, information stability, etc., are removed, do completely general formulas for entropy, divergence and mutual information exist ?

The answer is indeed in the affirmative for mutual information. In [10], Verdú and Han show that the channel capacity of arbitrary single-user channels is equal to the supremum, over all input processes, of the input-output *(mutual) inf-information rate*. By adopting the same technique as in [10], general expressions for the capacity of single-user channels with feedback and for Neyman-Pearson type-II error exponents are derived in [3] and [2], respectively. Furthermore, an application of the type-II error exponent formula to the non-feedback and feedback channel reliability functions is demonstrated in [2] and [5].

We therefore remark that the error probability of any stochastic system is actually characterized by a *sequence* of random variables. In the case of channel capacity, this sequence consists of the normalized information densities (evaluated under the optimal input process); while in the case of the Neyman-Pearson exponent, it consists of the normalized log-likelihood ratios of the null hypothesis distribution against the

*The work of P.-N. Chen is supported in part by ITRI. The work of F. Alajaji is supported in part by NSERC.

alternative hypothesis distribution (evaluated under the null hypothesis distribution). The ultimate CDF (cumulative distribution function) of this random sequence, which will be referred to in the next section as the *sup-spectrum* of the sequence, then determines the achievable error probability as illustrated in Figure 1.



x is: 1. general ε-capacity
2. general type-II error exponent
3. general source compression ratio

y is: 1. achievable channel coding error ε
2. achievable type-I
3. achievable probability of correctly decoding

Figure 1: The random sequences are respectively: the normalized information density under the optimal input process for case 1, the normalized log-likelihood ratio for case 2, and the normalized source density for case 3.

The same argument can also be applied to the source coding theorem; in this case, the decisive sequence of random variables is the normalized source density. The same figure can therefore be used to illustrate the relationship between the achievable source coding error and the (lossless) data compression ratio. These observations lead us to conclude that such methodology – which consists of using the spectrum of a decisive random sequence to formulate the corresponding error probability – is feasible for all three information quantities.

Another issue considered in this paper is to coincide the general expressions (which are written in terms of the spectrum of the corresponding decisive random sequence) with the conventionally known formulas. In [10], the general expression for ε-capacity is of the form

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

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

where $F_X(\cdot)$ is the sup-spectrum of the normalized information density. The general formula for the type-II error exponent derived in [2] has a similar form:

$$\sup\{D : \bar{F}(D) < \varepsilon\} \le \mathcal{B}(\varepsilon) \le \sup\{D : \bar{F}(D) \le \varepsilon\},$$

where $\bar{F}(D)$ is the sup-spectrum of the normalized log-likelihood ratio of the null hypothesis distribution against the alternative hypothesis distribution, and ε is the largest type-I error probability achievable subject to a lower bound on the type-II error exponent $\mathcal{B}$. In terms of notation, one may find some difficulty in connecting the above formulas to the conventional expressions of the same quantities; this may therefore obscure their physical meanings.

In this paper, we re-formulate the general expressions of entropy, mutual information, and divergence so that they coincide with their conventional counterparts as listed in Appendix A. The basic properties of these quantities are also analyzed. The relationships of these general expressions to the respective Shannon theorems are presented in Appendix B.

II. General formulas and properties

For a sequence of random variables $\{G_n\}_{n=1}^{\infty}$, the *inf-spectrum* $\underline{u}(\cdot)$ and *sup-spectrum* $\bar{u}(\cdot)$ are defined:

$$\underline{u}(\varepsilon) \triangleq \liminf_{n\to\infty} Pr\{G_n \le \varepsilon\},$$

$$\bar{u}(\varepsilon) \triangleq \limsup_{n\to\infty} Pr\{G_n \le \varepsilon\}.$$

Its *liminf in probability* is defined as the largest extended real number $\underline{U}$ such that for all $\xi > 0$,

$$\liminf_{n\to\infty} Pr\{G_n \le \underline{U} - \xi\} = \limsup_{n\to\infty} Pr\{G_n \le \underline{U} - \xi\} = 0,$$

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

$$\liminf_{n\to\infty} Pr\{G_n \ge \bar{U} + \xi\} = \limsup_{n\to\infty} Pr\{G_n \ge \bar{U} + \xi\} = 0.$$

Defining $\underline{U}_\delta$ and $\bar{U}_\delta$ by

$$\underline{U}_\delta \triangleq \sup\{\theta : \bar{u}(\theta) \le \delta\}, \text{ and } \bar{U}_\delta \triangleq \sup\{\theta : \underline{u}(\theta) \le \delta\},$$

respectively, it can clearly be concluded that $\underline{U} = \underline{U}_0$, and $\bar{U} = \bar{U}_{1^-}$, where the superscript "-" denotes a strict inequality in the definition of $\bar{U}_{1^-}$; i.e., $\bar{U}_{\delta^-} \triangleq \sup\{\theta : \underline{u}(\theta) < \delta\}$. For a better understanding of these quantities, we depict them in Figure 2. Based on the above notations, the general expressions of entropy, mutual information, and divergence are re-formulated in Appendix A. Some properties of entropy, mutual information, and divergence are listed in the next theorem [4].

Theorem *For $\delta, \gamma, \delta + \gamma \in [0, 1]$, the following statements hold.*

1. *$\bar{H}_\delta(\boldsymbol{X}) \ge 0$. $\bar{H}_\delta(\boldsymbol{X}) = 0$ if and only if $\{X^n = (X_1^{(n)}, \ldots, X_n^{(n)})\}_{n=1}^{\infty}$ is ultimately deterministic in probability. (also applies to $\underline{H}_\delta(\boldsymbol{X})$, $\bar{I}_\delta(\boldsymbol{X};\boldsymbol{Y})$, $\underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y})$, $\bar{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}})$, and $\underline{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}})$.)*
2. *$\underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) = \underline{I}_\delta(\boldsymbol{Y};\boldsymbol{X})$ and $\bar{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) = \bar{I}_\delta(\boldsymbol{Y};\boldsymbol{X})$.*

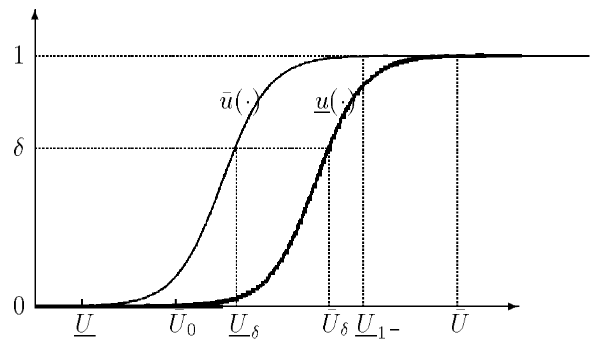


Figure 2: The ultimate CDF of a sequence of random variables $\{G_n\}_{n=1}^{\infty}$. $\bar{u}(\cdot)$ = sup-spectrum of G_n; $\underline{u}(\cdot)$ = inf-spectrum of G_n.

3.
$$\underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) \le \underline{H}_{\delta+\gamma}(\boldsymbol{Y}) - \underline{H}_{\gamma}(\boldsymbol{Y}|\boldsymbol{X}),$$
$$\underline{I}_{\delta+\gamma}(\boldsymbol{X};\boldsymbol{Y}) \ge \underline{H}_{\delta}(\boldsymbol{Y}) - \bar{H}_{(1-\gamma)^-}(\boldsymbol{Y}|\boldsymbol{X}).$$

4. $0 \le \underline{H}_\delta(\boldsymbol{X}) \le \bar{H}_\delta(\boldsymbol{X}) \le \log|\mathcal{X}|$, *where each* $X_i^{(n)} \in \mathcal{X}$, $i = 1,\ldots,n$ *and* $n = 1,2,\ldots$, *and* $\mathcal{X}$ *is finite.*

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

6. *(Data processing lemma) Suppose* X_1^n *and* X_3^n *are conditionally independent given* X_2^n. *Then* $\underline{I}_\delta(\boldsymbol{X}_1;\boldsymbol{X}_3) \le \underline{I}_\delta(\boldsymbol{X}_1;\boldsymbol{X}_2)$.

7. *(Optimality of independent inputs) Consider a finite alphabet, discrete memoryless channel – i.e.,* $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$, *for all* n. *For any input* $\boldsymbol{X}$ *and its corresponding output* $\boldsymbol{Y}$,
$$\underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) \le \underline{I}_\delta(\bar{\boldsymbol{X}};\bar{\boldsymbol{Y}}) = \underline{I}(\bar{\boldsymbol{X}};\bar{\boldsymbol{Y}}),$$
where $\bar{\boldsymbol{Y}}$ *is the output due to* $\bar{\boldsymbol{X}}$, *which is an independent process with the same first order statistics as* $\boldsymbol{X}$, *i.e.,* $P_{\overline{X}^n} = \prod_{i=1}^n P_{X_i}$.

In Appendix B, we generalize the Shannon coding theorems in terms of these new formulas. These general coding theorems lead us to make the following observations.

- The block source coding theorem in [8], which states that the minimum achievable fixed-length source coding rate of any finite-alphabet source is $\bar{H}(\boldsymbol{X}) = \bar{H}_{1^-}(\boldsymbol{X})$, has been generalized. Note that $\bar{H}(\boldsymbol{X})$ also denotes the general formula of the resolvability of $\boldsymbol{X}$, which represents the minimal number of random bits per sample required to reproduce the n-fold distribution of $\boldsymbol{X}$ with arbitrary accuracy as n grows to infinity [8].
- Consider the special case where $-(1/n)$ log $P_{X^n}(X^n)$ converges in probability to a constant H; this is indeed the AEP [6] which is a weaker condition than the convergence of $(1/n)\sum_{i=1}^n H(X_i)$, but implies information stability [9]. In this case, both $\underline{h}_X(\cdot)$ and $\bar{h}_X(\cdot)$ degenerate to a unit step function, yielding $\underline{H}(\boldsymbol{X}) = \bar{H}_\varepsilon(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}) = H$ for all $\varepsilon \in (0,1)$. Hence, our result reduces to the conventional source coding theorem [9, Theorem 1].
- More generally, if $-(1/n)\log P_{X^n}(X^n)$ converges in probability to a random variable Z whose CDF is $F_Z(\cdot)$, we have
$$Pe \approx 1 - F_Z(R) \quad \text{for} \quad R = \bar{H}_\varepsilon(\boldsymbol{X}) = \underline{H}_\varepsilon(\boldsymbol{X}).$$
Therefore, the relationship between the code rate and the ultimate optimal error probability is clearly defined as well.

Appendix A : General formulas for entropy, mutual-information, and divergence

Entropy

Observations : Arbitrary sequence of random source $\boldsymbol{X}$.

Decisive Random Sequence (normalized source density): $\frac{1}{n}h_{X^n}(X^n) \triangleq -\frac{1}{n}\log P_{X^n}(X^n)$.

Sup-Spectrum: $\bar{h}_X(\theta) \triangleq \limsup_{n\to\infty} Pr\left\{\frac{1}{n}h_{X^n}(X^n) \le \theta\right\}$.

Inf-Spectrum: $\underline{h}_X(\theta) \triangleq \liminf_{n\to\infty} Pr\left\{\frac{1}{n}h_{X^n}(X^n) \le \theta\right\}$.

General Formula: $\begin{cases} \underline{H}_\delta(\boldsymbol{X}) \triangleq \sup\{\theta : \bar{h}_X(\theta) \le \delta\} \\ \bar{H}_\delta(\boldsymbol{X}) \triangleq \sup\{\theta : \underline{h}_X(\theta) \le \delta\}. \end{cases}$

Sup-Entropy Rate : $\bar{H}(\boldsymbol{X}) \triangleq \bar{H}_{1^-}(\boldsymbol{X})$.

Inf-Entropy Rate : $\underline{H}(\boldsymbol{X}) \triangleq \underline{H}_0(\boldsymbol{X})$.

Mutual information

Observations : Arbitrary sequence of channel input and output processes $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Decisive Random Sequence (normalized information density):
$$\frac{1}{n}i_{(X^n,Y^n)}(X^n;Y^n) \triangleq \frac{1}{n}\log\frac{dP_{X^nY^n}}{d(P_{X^n}\times P_{Y^n})}(X^n,Y^n).$$

Sup-Spectrum :
$$\bar{i}_{(X,Y)}(\theta) \triangleq \limsup_{n\to\infty} Pr\left\{\frac{1}{n}i_{(X^n,Y^n)}(X^n;Y^n) \le \theta\right\}.$$

Inf-Spectrum :
$$\underline{i}_{(X,Y)}(\theta) \triangleq \liminf_{n\to\infty} Pr\left\{\frac{1}{n}i_{(X^n,Y^n)}(X^n;Y^n) \le \theta\right\}.$$

General Formula: $\begin{cases} \underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) \triangleq \sup\{\theta : \bar{i}_{(X,Y)}(\theta) \le \delta\} \\ \bar{I}_\delta(\boldsymbol{X};\boldsymbol{Y}) \triangleq \sup\{\theta : \underline{i}_{(X,Y)}(\theta) \le \delta\}. \end{cases}$

Sup-Information Rate: $\bar{I}(\boldsymbol{X};\boldsymbol{Y}) \triangleq \bar{I}_{1^-}(\boldsymbol{X};\boldsymbol{Y})$.

Inf-Information Rate: $\underline{I}(\boldsymbol{X};\boldsymbol{Y}) \triangleq \underline{I}_0(\boldsymbol{X};\boldsymbol{Y})$.

Divergence

Observations : Arbitrary sequence of two random observations $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$.

Decisive Random Sequence (normalized log-likelihood ratio): $\frac{1}{n} d_{X^n}(X^n \| \hat{X}^n) \triangleq \frac{1}{n} \log \frac{dP_{X^n}}{dP_{\hat{X}^n}}(X^n)$.

Sup-Spectrum :

$$\bar{d}_{X\|\hat{X}}(\theta) \triangleq \limsup_{n\to\infty} Pr\left\{\frac{1}{n} d_{X^n}(X^n\|\hat{X}^n) \le \theta\right\}.$$

Inf-Spectrum :

$$\underline{d}_{X\|\hat{X}}(\theta) \triangleq \liminf_{n\to\infty} Pr\left\{\frac{1}{n} d_{X^n}(X^n\|\hat{X}^n) \le \theta\right\}.$$

General Formula: $\begin{cases} \underline{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}}) \triangleq \sup\{\theta : \bar{d}_{X\|\hat{X}}(\theta) \le \delta\} \\ \bar{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}}) \triangleq \sup\{\theta : \underline{d}_{X\|\hat{X}}(\theta) \le \delta\}. \end{cases}$

Sup-Divergence Rate: $\bar{D}(\boldsymbol{X}) \triangleq \bar{D}_{1^-}(\boldsymbol{X}\|\hat{\boldsymbol{X}})$.

Inf-Divergence Rate: $\underline{D}(\boldsymbol{X}) \triangleq \underline{D}_0(\boldsymbol{X}\|\hat{\boldsymbol{X}})$.

Appendix B : General Shannon theorems

General lossless source coding theorem

Minimum source coding error: $1-\varepsilon$.

General source compression ratio achievable: $\bar{H}_\varepsilon(\boldsymbol{X})$.

General lossy source coding theorem

Source coding distortion constraint: ε-distortion $< D$.

General source compression ratio achievable:

$$\underline{R}_{1-\varepsilon}(D) \triangleq \min_{\{P_{Y|X}\ :\ \varepsilon-\text{distortion} \le D\}} \underline{I}(\boldsymbol{X};\boldsymbol{Y}).$$

For a definition of ε-distortion refer to Appendix C.

General channel coding theorem

Minimum channel coding error: ε.

General channel capacity: $\underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y})$.

General Neyman-Pearson type-II error exponent of fixed test level

Type-I error bound: ε.

General limsup of type-II error exponent: $\bar{D}_\varepsilon(\boldsymbol{X}\|\hat{\boldsymbol{X}})$.

General liminf of type-II error exponent: $\underline{D}_\varepsilon(\boldsymbol{X}\|\hat{\boldsymbol{X}})$.

General Neyman-Pearson type-II error exponent of exponential test level

Achievable type-I and II error exponent pair:

$$(\underline{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \bar{D}_\varepsilon(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}})).$$

Achievable type-I and II error exponent pair:

$$(\bar{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \underline{D}_\varepsilon(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}})),$$

where $\hat{\boldsymbol{X}}^{(s)}$ exhibits a tilted distribution defined in [4].

Appendix C : Definition of ε-distortion

Given a sequence of (arbitrary) distortion measures $\rho_n(\cdot,\cdot)$ for an arbitrary source $\boldsymbol{X}$, a sequence of data compression codes $\{f_n(\cdot)\}_{n=1}^\infty$ for source $\boldsymbol{X}$ is said to have ε-distortion less than D if

$$\bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X}, \boldsymbol{f}(\boldsymbol{X})) < D,$$

where

$$\boldsymbol{f}(\boldsymbol{X}) \triangleq \{f_n(X^n)\}_{n=1}^\infty,$$

$$\bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X}, \boldsymbol{f}(\boldsymbol{X})) \triangleq \sup\{\theta : \underline{\lambda}_{(X,f(X))}(\theta) \le 1-\varepsilon\},$$

and

$$\underline{\lambda}_{(X,f(X))}(\theta) \triangleq \liminf_{n\to\infty} Pr\left\{\frac{1}{n}\rho_n(X^n, f_n(X^n)) \le \theta\right\}.$$

References

[1] R. E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, Massachusetts, 1988.

[2] P.-N. Chen, "General formulas for the Neyman-Pearson type-II error exponent subject to *fixed* and *exponential* type-I error bounds," *IEEE Trans. Inform. Theory*, vol. IT-42, no. 1, pp. 316–323, January 1996.

[3] P.-N. Chen 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, pp. 544–549, March 1995.

[4] P.-N. Chen and F. Alajaji, "General formulas for entropy, divergence and mutual information," submitted to *IEEE Trans. Inform. Theory*, March 1996.

[5] P.-N. Chen 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] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[7] R. L. Dobrushin, "General formulation of Shannon's main theorm in information theory," *American Mathematical Society Translations*, vol. 33, pp. 323–438, AMS. Providence, RI, USA, 1963.

[8] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT-39, no. 3, pp. 752–772, May 1993.

[9] S. Vembu, S. Verdú and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, vol. IT-41, no. 1, pp. 44–54, January 1995.

[10] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. IT-40, no. 4, pp. 1147–1157, July 1994.