Generalized Source Coding Theorems and Hypothesis Testing: Part II – Operational Limits

Po-Ning Chen
Dept. of Communications Engineering
National Chiao Tung University
1001, Ta-Hsueh Road, Hsin Chu
Taiwan 30050, R.O.C.

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

Key Words: Shannon theory, AEP, source coding theorems, hypothesis testing, Neyman-Pearson error exponent

Abstract

In light of the information measures introduced in Part I, a generalized version of the Asymptotic Equipartition Property (AEP) is proved. General fixed-length data compaction and data compression (source coding) theorems for arbitrary finite-alphabet sources are also established. Finally, the general expression of the Neyman-Pearson type-II error exponent subject to upper bounds on the type-I error probability is examined.

I. Introduction

In Part I of this paper [5], generalized versions of the inf/sup-entropy/information/divergence rates of Han and Verdú were proposed and analyzed. Equipped with these information measures, we herein demonstrate a generalized Asymptotic Equipartition Property (AEP) Theorem and establish expressions for the infimum $(1-\varepsilon)$-achievable (fixed-length) coding rate of an arbitrary finite-alphabet source $\boldsymbol{X}$. These expressions turn out to be the *counterparts* of the ε-capacity formulas in [11, Theorem 6]. We also prove a general data compression theorem; this theorem extends a recent rate-distortion theorem [9, Theorem 10(a)] by Steinberg and Verdú (cf the remarks at the end of Sections II.1 and II.2).

The Neyman-Pearson hypothesis testing problem examined in [4] is revisited in light of the generalized divergence measures.

Since this work is a continuation of [5], we refer the reader to [5] for the technical definitions of the information measures used in this paper.

II. General Source Coding Theorems

The role of a source code is to represent the output of a source efficiently. This is achieved by introducing some controlled distortion into the source, hence reducing its intrinsic information content. There are two classes of source codes: data compaction codes and data compression codes [2]. The objective of both types of codes is to minimize the source description rate of the codes subject to a fidelity criterion constraint. In the case of data compaction, the fidelity criterion consists of the probability of decoding error Pe. If Pe is made arbitrarily small, we obtain a traditional error-free (or lossless) source coding system. Data compression codes are a larger class of codes in the sense that the fidelity criterion used in the coding scheme is a general distortion measure. We herein demonstrate data compaction and data compression theorems for arbitrary (not necessarily stationary ergodic, information stable, etc.) sources.

In this section, we assume that the source alphabet $\mathcal{X}$ is *finite* [1].

[1] Actually, the theorems in this section also apply for sources with countable alphabets. We assume finite alphabets in order to avoid uninteresting cases (such as $\bar{H}_\varepsilon(\boldsymbol{X}) = \infty$) that might arise with countable alphabets.

1. Data compaction coding theorem

Definition 2.1 (e.g. [2]) A block code for data compaction is a set $\mathscr{C}_n$ consisting of $M \stackrel{\triangle}{=} |\mathscr{C}_n|$ codewords of blocklength n:

$$\mathscr{C}_n \stackrel{\triangle}{=} \{c_1^n, c_2^n, \ldots, c_M^n\},$$

where each n-tuple $c_i^n \in \mathcal{X}^n$, $i = 1, 2, \ldots, M$.

Definition 2.2 Fix $1 \geq \varepsilon \geq 0$. R is a $(1-\varepsilon)$-achievable data compaction rate for a source $\boldsymbol{X}$ if there exists a sequence of data compaction codes $\mathscr{C}_n$ with

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n| = R,$$

and

$$\limsup_{n\to\infty} Pe(\mathscr{C}_n) \leq 1 - \varepsilon,$$

where $Pe(\mathscr{C}_n) \stackrel{\triangle}{=} Pr\left(X^n \notin \mathscr{C}_n\right)$ is the probability of decoding error.

The infimum $(1-\varepsilon)$-achievable data compaction rate for $\boldsymbol{X}$ is denoted by $T_{1-\varepsilon}(\boldsymbol{X})$.

For discrete memoryless sources, the data compaction theorem is proved by choosing the codebook $\mathscr{C}_n$ to be the *(weakly) typical set* [2] and applying the Asymptotic Equipartition Property (AEP) [3][2] which states that $(1/n)h_{X^n}(X^n)$ converges to $H(X)$ with probability one (and hence in probability). The AEP – which implies that the probability of the typical set is close to one for sufficiently large n – also holds for stationary ergodic sources [3]. It is however invalid for more general sources – e.g., nonstationary, nonergodic sources. We herein demonstrate a generalized AEP theorem.

Theorem 2.1 (Generalized AEP) Fix $1 > \varepsilon > 0$. Given an arbitrary source $\boldsymbol{X}$, define

$$\mathcal{T}_n[R] \stackrel{\triangle}{=} \left\{x^n \in \mathcal{X}^n : -\frac{1}{n} \log P_{X^n}(x^n) \leq R\right\}.$$

Then $(\forall\ \gamma > 0)$ the following statements hold.

1.

$$\liminf_{n\to\infty} Pr\left\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right\} \leq \varepsilon \tag{2.1}$$

2.

$$\liminf_{n\to\infty} Pr\left\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]\right\} > \varepsilon \tag{2.2}$$

3. The number of elements in $\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})]$, denoted by $\left|\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})]\right|$, satisfies

$$\left|\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]-\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right| \leq \exp\left\{n(\bar{H}_\varepsilon(\boldsymbol{X})+\gamma)\right\}. \tag{2.3}$$

4. $(\forall\ \gamma>0)(\exists\ \rho=\rho(\gamma)>0, N_0$ and a subsequence $\{n_j\}_{j=1}^{n}$ such that $\forall n_j > N_0)$,

$$\left|\mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]-\mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right| > \rho(\gamma)\exp\left\{n_j(\bar{H}_\varepsilon(\boldsymbol{X})-\gamma)\right\}, \tag{2.4}$$

where the operation $A-B$ between two sets A and B is defined by $A-B\triangleq A\cap B^c$, with B^c denoting the complement set of B.

***Proof*:** (2.1) and (2.2) follows from the definitions. For (2.3), we have

$$\begin{aligned}
1 &\geq \sum_{x^n\in\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]-\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]} P_{X^n}(x^n)\\
&\geq \sum_{x^n\in\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]-\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]} \exp\left\{-n\left(\bar{H}_\varepsilon(\boldsymbol{X})+\gamma\right)\right\}\\
&= \left|\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]-\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right| \exp\left\{-n\left(\bar{H}_\varepsilon(\boldsymbol{X})+\gamma\right)\right\}.
\end{aligned}$$

It remains to show (2.4). (2.2) implies that there exist $\rho=\rho(\gamma)>0$ and N_1 such that for all $n>N_1$,

$$Pr\left\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]\right\} > \varepsilon+2\rho(\gamma).$$

Furthermore, (2.1) implies that for the previously chosen $\rho(\gamma)$, there exist N_2 and a subsequence $\{n_j\}_{j=1}^{\infty}$ such that for all $n_j>N_2$,

$$Pr\left\{\mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right\} < \varepsilon+\rho(\gamma).$$

Therefore, for all $n_j > N_0 \stackrel{\triangle}{=} \max(N_1, N_2)$,

$$\begin{aligned}\rho(\gamma) &< Pr\left\{\mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma] - \mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right\} \\ &< \left|\mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma] - \mathcal{T}_{n_j}[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma]\right| \exp\left\{-n_j\left(\bar{H}_\varepsilon(\boldsymbol{X})-\gamma\right)\right\}.\end{aligned}$$

□

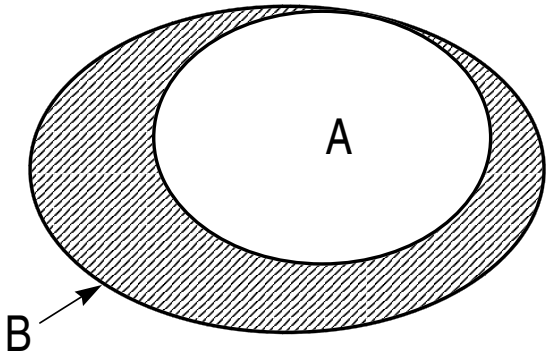

Figure 1: Illustration of the Generalized AEP Theorem. $A = \mathcal{T}_n[\bar{H}_\varepsilon(\mathbf{X}) - \gamma]$, $B = \mathcal{T}_n[\bar{H}_\varepsilon(\mathbf{X}) + \gamma]$, and $B - A =$ dashed region.

Comment: With the illustration depicted in Figure 1, we can clearly deduce that Theorem 2.1 is *indeed* a generalized version of the AEP since:

- The set

  $$B - A \stackrel{\triangle}{=} \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma] - \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})-\gamma] = \left\{x^n \in \mathcal{X}^n : \left|-\frac{1}{n}\log P_{X^n}(x^n) - \bar{H}_\varepsilon(\boldsymbol{X})\right| \leq \gamma\right\}$$

  is nothing but the typical set.

- (2.1) and (2.2) $\Rightarrow$ that $q \stackrel{\triangle}{=} Pr(B - A) > 0$ infinitely often.

- (2.3) and (2.4) $\Rightarrow$ that the number of sequences in $B - A$ (the dashed region) is approximatively equal to $\exp\left\{n\bar{H}_\varepsilon(\boldsymbol{X})\right\}$, and the probability of each sequence in $B - A$ is $\approx q \times \exp\left\{-n\bar{H}_\varepsilon(\boldsymbol{X})\right\}$.

- In particular, if $\boldsymbol{X}$ is a stationary ergodic source, then $\bar{H}_\varepsilon(\boldsymbol{X})$ is independent of ε and $\bar{H}_\varepsilon(\boldsymbol{X}) = \underline{H}_\varepsilon(\boldsymbol{X}) = H \ \forall\ \varepsilon \in (0,1)$, where H is the source entropy rate

$$H = \lim_{n\to\infty} \frac{1}{n} E_{P_{X^n}} \left[-\log P_{X^n}(X^n) \right].$$

  In this case, (2.1)-(2.2) and the fact that $\bar{H}_\varepsilon(\boldsymbol{X}) = \underline{H}_\varepsilon(\boldsymbol{X})\ \forall \varepsilon$ imply that the probability q of the typical set $B - A$ is close to one (for n sufficiently large), and (2.3) and (2.4) imply that there are about e^{nH} typical sequences of length n, each with probability about e^{-nH}. Hence we obtain the conventional AEP (cf [3, Theorem 3.1.2] or [2, Theorem 3.4.2]).

We now apply Theorem 2.1 to prove a *general* data compaction theorem for block codes.

Theorem 2.2 (General data compaction theorem) Fix $1 > \varepsilon > 0$. For any source $\boldsymbol{X}$,

$$\bar{H}_{\varepsilon^-}(\boldsymbol{X}) \leq T_{1-\varepsilon}(\boldsymbol{X}) \leq \bar{H}_\varepsilon(\boldsymbol{X}).$$

Note that actually $T_{1-\varepsilon}(\boldsymbol{X}) = \bar{H}_{\varepsilon^-}(\boldsymbol{X})$, since $T_{1-\varepsilon}(\boldsymbol{X})$ is left-continuous in ε (cf Appendix B).

***Proof*:**

1. *Forward part (achievability):* We need to prove the existence of a sequence of block codes $\mathscr{C}_n$ with

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n| < \bar{H}_\varepsilon(\boldsymbol{X}) + 2\gamma,$$

and

$$\limsup_{n\to\infty} Pe(\mathscr{C}_n) \leq 1 - \varepsilon.$$

Choose the code to be $\mathscr{C}_n = \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) + \gamma]$. Then by definition of $\mathcal{T}_n[\cdot]$,

$$|\mathscr{C}_n| = \left| \mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) + \gamma] \right| \leq \exp\left\{ n \left(\bar{H}_\varepsilon(\boldsymbol{X}) + \gamma \right) \right\}.$$

Therefore

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n| \leq \bar{H}_\varepsilon(\boldsymbol{X}) + \gamma < \bar{H}_\varepsilon(\boldsymbol{X}) + 2\gamma.$$

On the other hand,

$$1 - Pe(\mathscr{C}_n) = Pr\{\mathscr{C}_n\} = Pr\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X}) + \gamma]\},$$

which implies from (2.2) that

$$\liminf_{n\to\infty}[1-Pe(\mathscr{C}_n)]=\liminf_{n\to\infty}Pr\{\mathcal{T}_n[\bar{H}_\varepsilon(\boldsymbol{X})+\gamma]\}>\varepsilon.$$

Hence,

$$\limsup_{n\to\infty}Pe(\mathscr{C}_n)<1-\varepsilon.$$

Accordingly, $T_{1-\varepsilon}<\bar{H}_\varepsilon(\boldsymbol{X})+2\gamma$ for any $\gamma>0$. This proves the forward part.

To show the converse part, we need the following remark.

Remark: For all $x^n\in\mathscr{C}_n^*$ and $\hat{x}^n\notin\mathscr{C}_n^*$,

$$P_{X^n}(x^n)\ \geq\ P_{X^n}(\hat{x}^n),$$

where $\mathscr{C}_n^*$ is the optimal block code defined as follows: for any block code $\mathscr{C}_n$ with $|\mathscr{C}_n|\ =\ |\mathscr{C}_n^*|$, $Pe(\mathscr{C}_n^*)\ \leq\ Pe(\mathscr{C}_n)$.

This result follows directly from the definition of $\mathscr{C}_n^*$ and the fact that $Pe(\mathscr{C}_n^*)=P_{X^n}([\mathscr{C}_n^*]^c)$. The above remark indeed points out that the optimal code must be of the shape

$$\left\{x^n\in\mathcal{X}^n:-\frac{1}{n}\log P_{X^n}(x^n)<R\right\}\subset\mathscr{C}_n^*\subset\left\{x^n\in\mathcal{X}^n:-\frac{1}{n}\log P_{X^n}(x^n)\leq R\right\}. \tag{2.5}$$

2. *Converse part:* We show that for all codes with code rate

$$R\stackrel{\triangle}{=}\limsup_{n\to\infty}(1/n)\log|\mathscr{C}_n|<\bar{H}_{\varepsilon^-}(\boldsymbol{X}),$$

$$\limsup_{n\to\infty}Pe(\mathscr{C}_n)>1-\varepsilon.$$

By definition of $\bar{H}_{\varepsilon^-}(\boldsymbol{X})$, there exists $0<\varepsilon'<\varepsilon$ such that $R<\bar{H}_{\varepsilon'}(\boldsymbol{X})\leq\bar{H}_{\varepsilon^-}(\boldsymbol{X})$. Since $Pe(\mathscr{C}_n^*)\leq Pe(\mathscr{C}_n)$ for $\mathscr{C}_n^*$ with the same size as $\mathscr{C}_n$, we only need to show

$$\limsup_{n\to\infty}Pe(\mathscr{C}_n^*)>1-\varepsilon.$$

(2.5) already gives us the shape of the optimal block code. We claim that the set $\mathcal{T}_n[\bar{H}_{\varepsilon'}(\boldsymbol{X})+\gamma]-\mathcal{T}_n[\bar{H}_{\varepsilon'}(\boldsymbol{X})]$ is not contained in $\mathscr{C}_n^*$ for any $\gamma>0$ infinitely often because if it were, then by slightly modifying the proof of (2.4), it can be shown that there exists $\gamma>0$ such that

$$\left|\mathscr{C}_{n_j}^*\right|>\left|\mathcal{T}_{n_j}[\bar{H}_{\varepsilon'}(\boldsymbol{X})+\gamma]-\mathcal{T}_{n_j}[\bar{H}_{\varepsilon'}(\boldsymbol{X})]\right|>\rho(\gamma)\exp\left\{n_j\left(\bar{H}_{\varepsilon'}(\boldsymbol{X})\right)\right\}$$

for some positive $\rho(\gamma)$, subsequence $\{n_j\}_{j=1}^{\infty}$ and sufficiently large j, implying that

$$R \geq \bar{H}_{\varepsilon'}(\boldsymbol{X}). \tag{2.6}$$

This violates the code rate constraint $R < \bar{H}_{\varepsilon'}(\boldsymbol{X})$. Hence, $\mathscr{C}_n^*$ is a subset of $\mathcal{T}_n[\bar{H}_{\varepsilon'}(\boldsymbol{X})]$ for all but finitely many n. Consequently,

$$\liminf_{n\to\infty}[1 - Pe(\mathscr{C}_n^*)] = \liminf_{n\to\infty} Pr(\mathscr{C}_n^*) \leq \liminf_{n\to\infty} Pr\{\mathcal{T}_n[\bar{H}_{\varepsilon'}(\boldsymbol{X})]\} \leq \varepsilon' < \varepsilon,$$

where the last inequality follows from the definition of $\bar{H}_{\varepsilon'}(\boldsymbol{X})$. This immediately implies that

$$\limsup_{n\to\infty} Pe(\mathscr{C}_n^*) > 1 - \varepsilon.$$

This proves the converse part. □

Observations:

- For the sake of clarity, we only considered in Theorem 2.2 the case where $\varepsilon \in (0, 1)$. We can however easily extend the result to the cases where $\varepsilon = 0$ and $\varepsilon = 1$. By definition, $\bar{H}_{0^-}(\boldsymbol{X}) = -\infty$ and $\bar{H}_1(\boldsymbol{X}) = \infty$. Therefore, to show that Theorem 2.2 holds for $\varepsilon = 0$ and $\varepsilon = 1$, it suffices to prove that

  $$T_1(\boldsymbol{X}) \leq \bar{H}_0(\boldsymbol{X}) \tag{2.7}$$

  and

  $$T_0(\boldsymbol{X}) \geq \bar{H}_{1^-}(\boldsymbol{X}). \tag{2.8}$$

  The validity of (2.7) follows form the proof of the forward-part of Theorem 2.2 ; similarly, (2.8) can be verified using the same arguments in the proof of the converse-part of Theorem 2.2 .

- Theorem 2.2 is indeed the *counterpart* of the result on the channel ε-capacity in [11, Theorem 6]. It describes, in terms of the parameter ε, the relationship between the code rate and the ultimate probability of decoding error:

  $$Pe \approx 1 - \varepsilon \quad \text{and} \quad R = \bar{H}_{\varepsilon^-}(\boldsymbol{X}).$$

- Note that as $\varepsilon \uparrow 1$, $\bar{H}_{\varepsilon^-}(\boldsymbol{X}) \to \bar{H}_{1^-}(\boldsymbol{X}) = \bar{H}(\boldsymbol{X})$. Hence, this theorem generalizes the block source coding theorem in [8], which states that the minimum achievable fixed-length source coding rate of any finite-alphabet source is $\bar{H}(\boldsymbol{X})$.

- Consider the special case where $-(1/n)\log P_{X^n}(X^n)$ converges in probability to a constant H; this reduces Theorem 2.1 to the conventional AEP [3]. In this case, both $\underline{h}_{\boldsymbol{X}}(\cdot)$ and $\bar{h}_{\boldsymbol{X}}(\cdot)$ degenerate to a unit step function:

$$u(\theta - H) = \begin{cases} 1, & \text{if } \theta \geq H; \\ 0, & \text{if } \theta < H, \end{cases}$$

  yielding $\underline{H}(\boldsymbol{X}) = \bar{H}_{\varepsilon^-}(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}) = H$ for all $\varepsilon \in (0,1)$, where H is the source entropy rate. Hence, our result reduces to the conventional source coding theorem for information stable sources [10, Theorem 1].

- More generally, if $-(1/n)\log P_{X^n}(X^n)$ converges in probability to a random variable Z whose cumulative distribution function (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 also clearly defined.

  Example: Consider a binary exchangeable (hence stationary but nonergodic in general [1]) source $\boldsymbol{X}$. Then there exists a distribution G concentrated on the interval $(0,1)$ such that the process $\boldsymbol{X}$ is a mixture of Bernoulli(θ) processes where the parameter $\theta \in \Theta = (0,1)$ and has distribution G [1, Corollary 1]. In this case, it can be shown via the ergodic decomposition theorem that $-(1/n)\log P_{X^n}(X^n)$ converges in probability to $Z = h_b(\theta)$ [1][7], where $h_b(x) \stackrel{\triangle}{=} -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function. We therefore obtain that the cdf of Z is $F_Z(z) = P(h_b(\theta) \leq z)$ where θ has distribution G. Finally, note that as $\epsilon \uparrow 1$, $Pe \to 0$ and

$$\lim_{\epsilon \uparrow 1} \bar{H}_{\varepsilon^-}(\boldsymbol{X}) = \inf\ [r : dG(h_b(\theta) \leq r) = 1] \stackrel{\triangle}{=} \operatorname{ess}_{\Theta} \sup\ h_b(\theta).$$

  The above equation is indeed the minimum achievable (i.e., with $Pe \to 0$) fixed-length source coding rate for stationary nonergodic sources [6].

Remark: In this work, the definition that we adopt for the $(1-\varepsilon)$-achievable data compaction rate, is slightly different from the one used in [8, Definition 8]. As a result, our $T_{1-\varepsilon}(\boldsymbol{X})$ is right-continuous with respect to $(1-\varepsilon)$, and is equal to $\bar{H}_{\varepsilon^-}(\boldsymbol{X})$ for $\varepsilon \in (0,1]$ and 0 for $\varepsilon = 0$ (cf Appendix B); In fact, the definition in [8] also yields the same result, which was separately proved by Steinberg and Verdú as a direct consequence of Theorem 10(a) [9] (cf Corollary 3 in [9]). To be precise, their $T_{1-\varepsilon}(\boldsymbol{X})$, denoted by $T_e(1-\varepsilon, \boldsymbol{X})$ in [9], is shown for $0 < \varepsilon < 1$ to be equal to

$$\begin{aligned}
T_e(1-\varepsilon, \boldsymbol{X}) &= \bar{R}_v(2(1-\varepsilon)), (\text{cf Definition 17 in [9]}) \\
&= \inf\left\{\theta : \limsup_{n\to\infty} P_{X^n}\left[-\frac{1}{n}\log P_{X^n}(X^n) > \theta\right] \le 1-\varepsilon\right\} \\
&= \inf\left\{\theta : \liminf_{n\to\infty} P_{X^n}\left[-\frac{1}{n}\log P_{X^n}(X^n) \le \theta\right] \ge \varepsilon\right\} \\
&= \inf\{\theta : \underline{h}(\theta) \ge \varepsilon\} \\
&= \sup\{\theta : \underline{h}(\theta) < \varepsilon\} \\
&= \bar{H}_{\varepsilon^-}(\boldsymbol{X}).
\end{aligned}$$

Note that Theorem 10(a) in [9] is a data compression theorem for arbitrary sources which the authors show as a by-product of their results on finite-precision resolvability theory [9]. Here, we establish Theorem 2.2 in a different and more direct way; it is proven using the generalized entropy measure introduced in [5] and the Generalized AEP (Theorem 2.1). In the next section, we generalize Theorem 10(a) of [9].

2. Data compression coding theorem

Definition 2.3 (e.g. [2]) Given a source alphabet $\mathcal{X}$ and a reproduction alphabet $\mathcal{Y}$, a block code for data compression of blocklength n and size M is a mapping $f_n(\cdot) : \mathcal{X}^n \to \mathcal{Y}^n$ that results in $\|f_n\| = M$ codewords of length n, where each codeword is a sequence of n reproduction letters.

Definition 2.4 A distortion measure $\rho_n(\cdot,\cdot)$ is a mapping

$$\rho_n : \mathcal{X}^n \times \mathcal{Y}^n \to \Re^+ \stackrel{\triangle}{=} [0,\infty).$$

We can view the distortion measure as the cost of representing a source n-tuple X^n by a reproduction n-tuple $f_n(X^n)$.

In Theorem 10.(a) of [9], Steinberg and Verdú provide a data compression theorem for arbitrary sources under the restriction that the probability of excessive distortion due to the achievable data compression codes is equal to zero (cf Definitions 30 and 31 in [9]). We herein provide a generalization of their result by relaxing the restriction on the probability of excessive distortion.

Definition 2.5 (Distortion inf-spectrum and ε-sup-distortion rate) Let $\boldsymbol{X}$ and $\{\rho_n(\cdot,\cdot)\}_{n\geq 1}$ be given. Let $\boldsymbol{f}(\boldsymbol{X}) \stackrel{\triangle}{=} \{f_n(X^n)\}_{n=1}^{\infty}$ denote a sequence of data compression codes for $\boldsymbol{X}$. The *distortion inf-spectrum* $\underline{\lambda}_{(\boldsymbol{X},\boldsymbol{f}(\boldsymbol{X}))}(\theta)$ for $\boldsymbol{f}(\boldsymbol{X})$ is defined by

$$\underline{\lambda}_{(\boldsymbol{X},\boldsymbol{f}(\boldsymbol{X}))}(\theta) \stackrel{\triangle}{=} \liminf_{n\to\infty} Pr\left\{\frac{1}{n}\rho_n(X^n, f_n(X^n)) \leq \theta\right\}.$$

For any $1 > \varepsilon > 0$, the *ε-sup-distortion rate* $\bar{\Lambda}_\varepsilon(\boldsymbol{X}, \boldsymbol{f}(\boldsymbol{X}))$ is defined by

$$\bar{\Lambda}_\varepsilon(\boldsymbol{X}, \boldsymbol{f}(\boldsymbol{X})) \stackrel{\triangle}{=} \sup\{\theta : \underline{\lambda}_{(\boldsymbol{X},\boldsymbol{f}(\boldsymbol{X}))}(\theta) \leq \varepsilon\},$$

which is exactly the quantile of $\underline{\lambda}_{(\boldsymbol{X},\boldsymbol{f}(\boldsymbol{X}))}(\theta)$.

Definition 2.6 Fix $D > 0$ and $1 > \varepsilon > 0$. R is a $(1-\varepsilon)$-achievable data compression rate at distortion D for a source $\boldsymbol{X}$ if there exists a sequence of data compression codes $f_n(\cdot)$ with

$$\limsup_{n\to\infty} \frac{1}{n} \log \|f_n\| = R,$$

and $(1-\varepsilon)$-sup-distortion rate less than or equal to D:

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

Note that stating that the code has $(1-\varepsilon)$-sup-distortion rate less than or equal to D is equivalent to stating that the limsup of the probability of excessive distortion (i.e., distortion larger than D) is smaller than ε: $\limsup_{n\to\infty} Pr\left\{\frac{1}{n}\rho_n(X^n, f_n(X^n)) > D\right\} < \varepsilon$.

The infimum $(1-\varepsilon)$-achievable data compression rate at distortion D for $\boldsymbol{X}$ is denoted by $T_{1-\varepsilon}(D, \boldsymbol{X})$.

Theorem 2.3 (General data compression theorem) Fix $D > 0$ and $1 > \varepsilon > 0$. Let $\boldsymbol{X}$ and $\{\rho_n(\cdot,\cdot)\}_{n\geq 1}$ be given. Then

$$R_{(1-\varepsilon)^-}(D) \leq T_{1-\varepsilon}(D,\boldsymbol{X}) \leq R_{1-\varepsilon}(D),$$

where

$$R_{1-\varepsilon}(D) \stackrel{\triangle}{=} \inf_{\{P_{\boldsymbol{Y}|\boldsymbol{X}} : \bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X},\boldsymbol{Y}) \leq D\}} \bar{I}(\boldsymbol{X};\boldsymbol{Y}),$$

and $P_{\boldsymbol{Y}|\boldsymbol{X}} = \{P_{Y^n|X^n}\}_{n=1}^{\infty}$ denotes a sequence of conditional distributions satisfying the constraint $\bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X},\boldsymbol{Y})) \leq D$.

In other words, $T_{1-\varepsilon}(D,\boldsymbol{X}) = R_{1-\varepsilon}(D)$, except possibly at the points of discontinuities of $R_{1-\varepsilon}(D)$ (which are countable).

***Proof*:**

1. *Forward part (achievability):* Choose $\gamma > 0$. We will prove the existence of a sequence of block codes with

$$\limsup(1/n)\log|\mathcal{C}_n| < R_{1-\varepsilon}(D) + 2\gamma,$$

and

$$\bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X}; f(\boldsymbol{X})) < D + \gamma.$$

step 1: Let $P_{\bar{\boldsymbol{Y}}|\boldsymbol{X}}$ be the distribution achieving $R_{1-\varepsilon}(D)$, and let $P_{\bar{\boldsymbol{Y}}}$ be the $\boldsymbol{Y}$-marginal of $P_{\boldsymbol{X}}P_{\bar{\boldsymbol{Y}}|\boldsymbol{X}}$.

step 2: Let R satisfy $R_{1-\varepsilon}(D)+2\gamma > R > R_{1-\varepsilon}(D)+\gamma$. Choose $M = e^{nR}$ n-blocks independently according to $P_{\bar{\boldsymbol{Y}}}$, and denote the resulting random set by $\mathcal{C}_n$.

step 3: For a given $\mathcal{C}_n$, we denote by $A(\mathcal{C}_n)$ the set of sequences $x^n \in \mathcal{X}^n$ such that there exists $y^n \in \mathcal{C}_n$ with

$$\frac{1}{n}\rho_n(x^n,y^n) \leq D + \gamma.$$

step 4: *Claim:*

$$\limsup_{n\to\infty} E_{\bar{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n))] < \varepsilon.$$

The proof of this claim is provided in Appendix A.

Therefore there exists (a sequence of) $\mathcal{C}_n^*$ such that

$$\limsup_{n\to\infty} P_{X^n}(A^c(\mathcal{C}_n^*)) < \varepsilon.$$

step 5: Define a sequence of codes $\{f_n\}$ by

$$f_n(x^n) = \begin{cases} \arg \min\limits_{y^n \in \mathcal{C}_n^*} \rho_n(x^n, y^n), & \text{if } x^n \in A(\mathcal{C}_n^*); \\ \underline{0}, & \text{otherwise,} \end{cases}$$

where $\underline{0}$ is a fixed default n-tuple in $\mathcal{Y}^n$.

Then

$$\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\rho_n(x^n, f_n(x^n)) \le D + \gamma\right\} \supset A(\mathcal{C}_n^*),$$

since $(\forall\ x^n \in A(\mathcal{C}_n^*))$ there exists $y^n \in \mathcal{C}_n^*$ such that $(1/n)\rho_n(x^n, y^n) \le D + \gamma$, which by definition of f_n implies that $(1/n)\rho_n(x^n, f_n(x^n)) \le D + \gamma$.

step 6: Consequently,

$$\begin{aligned} \underline{\lambda}_{(\boldsymbol{X}, f(\boldsymbol{X}))}(D+\gamma) &= \liminf_{n\to\infty} P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\rho_n(x^n, f(x^n)) \le D + \gamma\right\} \\ &\ge \liminf_{n\to\infty} P_{X^n}(A(\mathcal{C}_n^*)) \\ &= 1 - \limsup_{n\to\infty} P_{X^n}(A^c(\mathcal{C}_n^*)) \\ &> 1 - \varepsilon. \end{aligned}$$

Hence,

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

where the last step is clearly depicted in Figure 2.

This proves the forward part.

2. *Converse part:* We show that for any sequence of encoders $\{f_n(\cdot)\}_{n=1}^{\infty}$, if

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

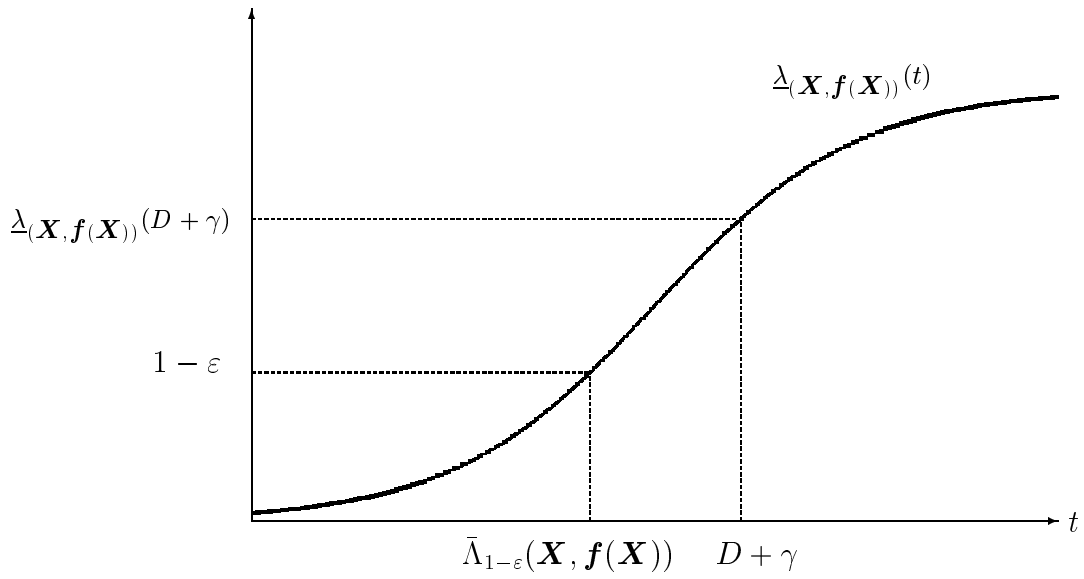

Figure 2: $\underline{\lambda}_{(\mathbf{X},f(\mathbf{X}))}(D+\gamma) > 1-\varepsilon \Rightarrow \bar{\Lambda}_{1-\varepsilon}(\mathbf{X}, f(\mathbf{X})) < D+\gamma$.

then

$$\limsup_{n\to\infty} \frac{1}{n} \log \|f_n\| \geq R_{(1-\varepsilon)^-}(D).$$

Let

$$P_{\hat{Y}^n|X^n}(y^n|x^n) \stackrel{\triangle}{=} \begin{cases} 1, & \text{if } y^n = f_n(x^n); \\ 0, & \text{otherwise.} \end{cases}$$

Then to evaluate the statistical properties of the random variable $(1/n)\rho_n(X^n, f_n(X^n))$ under distribution P_{X^n} is equivalent to evaluating the random variable $(1/n)\rho_n(X^n, \hat{Y}^n)$ under distribution $P_{X^n\hat{Y}^n}$. Therefore

$$\begin{aligned} R_{(1-\varepsilon)^-}(D) &\stackrel{\triangle}{=} \inf_{\{P_{\boldsymbol{Y}|\boldsymbol{X}} : \bar{\Lambda}_{(1-\varepsilon)^-}(\boldsymbol{X},\boldsymbol{Y}) \leq D\}} \bar{I}(\boldsymbol{X};\boldsymbol{Y}) \\ &\leq \bar{I}(\boldsymbol{X};\hat{\boldsymbol{Y}}) \\ &\leq \bar{H}(\hat{\boldsymbol{Y}}) - \underline{H}(\hat{\boldsymbol{Y}}|\boldsymbol{X}) \\ &\leq \bar{H}(\hat{\boldsymbol{Y}}) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \|f_n\|, \end{aligned}$$

where the second inequality follows from [5, Lemma 3.2] (cf (3.12) with $\gamma = 1^-$ and $\delta = 0$), and the third inequality follows from the fact that $\underline{H}(\hat{\boldsymbol{Y}}|\boldsymbol{X}) \geq 0$.

□

Observations:

1. *Comparison with Steinberg and Verdú's result* [9]. If $\varepsilon \downarrow 0$, then

$$R_{1-\varepsilon}(D) \uparrow R_{1^-}(D) \stackrel{\triangle}{=} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}:\bar{\Lambda}_{1^-}(\boldsymbol{X},\boldsymbol{Y})\leq D} \bar{I}(\boldsymbol{X};\boldsymbol{Y}).$$

   Remark that $R_{1^-}(D)$ is nothing but the *sup rate-distortion function* $\bar{\boldsymbol{R}}(D)$ described in Definition 14 of [9]. Therefore, this theorem reduces to Theorem 10.(a) of [9] when $\varepsilon \downarrow 0$. Note that according to the terminology of [9, Definition 14], $R_{1-\varepsilon}(D)$ may be called the $(1-\varepsilon)$-*sup rate-distortion function*.

2. *Comparison with the data compaction theorem.* For the probability-of-error distortion measure $\rho_n : \mathcal{X}^n \to \mathcal{X}^n$, namely,

$$\rho_n(x^n, \hat{x}^n) = \begin{cases} n, & \text{if } x^n \neq \hat{x}^n; \\ 0, & \text{otherwise,} \end{cases}$$

   we define a data compression code $f_n : \mathcal{X}^n \to \mathcal{X}^n$ based on a chosen data compaction code book $\mathcal{C}_n \subset \mathcal{X}^n$:

$$f_n(x^n) = \begin{cases} x^n, & \text{if } x^n \in \mathcal{C}_n; \\ \underline{0}, & \text{if } x^n \notin \mathcal{C}_n, \end{cases}$$

   where $\underline{0}$ is some default element in $\mathcal{X}^n$. Then $(1/n)\rho_n(x^n, f_n(x^n))$ is either 1 or 0 which results in a cumulative distribution function as shown in Figure 3. Consequently, for any $\delta \in [0, 1)$,

$$Pr\left\{\frac{1}{n}\rho_n(X^n, f_n(X^n)) \leq \delta\right\} \;=\; Pr\left\{X^n = f_n(X^n)\right\}.$$

   In other words, the condition

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

   is equivalent to

$$\liminf_{n\to\infty} Pr\{X^n = f_n(X^n)\} > 1 - \varepsilon,$$

   which is exactly the same as $\limsup_{n\to\infty} Pr\{X^n \neq f_n(X^n)\} < \varepsilon$.

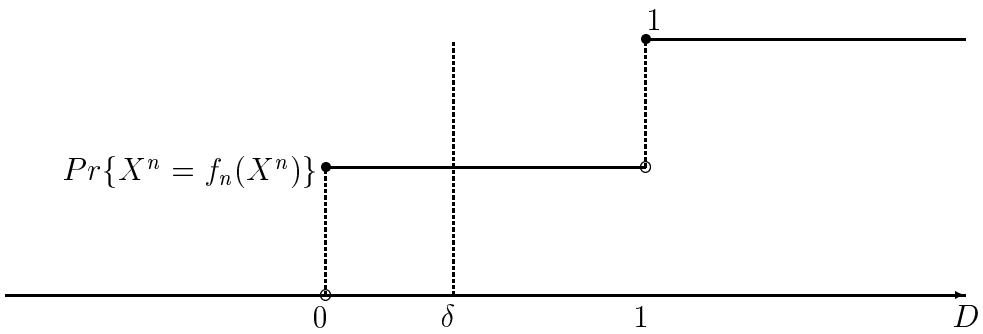

Figure 3: The CDF of $(1/n)\rho_n(X^n, f_n(X^n))$ for the probability-of-error distortion measure.

By comparing the source compaction and compression theorems, we remark that $\bar{H}_{1-\varepsilon}(\boldsymbol{X})$ is indeed the counterpart of $R_{1-\varepsilon}(\delta)$ for the probability-of-error distortion measure and $\delta \in [0, 1)$. In particular, in the extreme case where ε goes to zero,

$$\bar{H}(\boldsymbol{X}) = \inf_{\{P_{\hat{\boldsymbol{X}}|\boldsymbol{X}} : \limsup_{n\to\infty} Pr(X^n \neq \hat{X}^n) = 0\}} \bar{I}(\boldsymbol{X}; \hat{\boldsymbol{X}}),$$

which follows from the fact that (cf (3.12) and (3.14) in [5, Lemma 3.2])

$$\bar{H}(\boldsymbol{X}) - \bar{H}(\boldsymbol{X}|\hat{\boldsymbol{X}}) \leq \bar{I}(\boldsymbol{X}; \hat{\boldsymbol{X}}) \leq \bar{H}(\boldsymbol{X}) - \underline{H}(\boldsymbol{X}|\hat{\boldsymbol{X}}),$$

and $\bar{H}(\boldsymbol{X}|\hat{\boldsymbol{X}}) = \underline{H}(\boldsymbol{X}|\hat{\boldsymbol{X}}) = 0$. Therefore, in this case, the data compression theorem reduces (as expected) to the data compaction theorem (Theorem 2.2).

III. Neyman-Pearson Hypothesis Testing

In Neyman-Pearson hypothesis testing, the objective is to decide between two different explanations for the observed data. More specifically, given a sequence of observations with unknown underlying distribution Q, we consider two hypotheses:

- $H_0: \quad Q = P_{\boldsymbol{X}}$ (null hypothesis).
- $H_1: \quad Q = P_{\hat{\boldsymbol{X}}}$ (alternative hypothesis).

If we accept hypothesis H_1 when H_0 is actually true, we obtain what is known as a *type-I error*, and the probability of this event is denoted by α [2]. Accepting hypothesis H_0 when H_1 is actually true results in what we call a *type-II error*; the probability of this event is denoted by β. In general, the goal is to minimize *both* error probabilities; but there is a tradeoff since if α is reduced beyond a certain threshold then β increases (and vice-versa). Hence, we minimize one of the error probabilities subject to a constraint on the other error probability.

In the case of an arbitrary sequence of observations, the general expression of the Neyman-Pearson type-II error exponent subject to a constant bound on the type-I error has been proved in [4, Theorem 1]. We re-formulate the expression in terms of the ε-inf/sup-divergence rates in the next theorem.

Theorem 3.4 (Neyman-Pearson type-II error exponent for a fixed test level)

Consider a sequence of random observations which is assumed to have a probability distribution governed by either $P_{\boldsymbol{X}}$ (null hypothesis) or $P_{\hat{\boldsymbol{X}}}$ (alternative hypothesis). Then, the type-II error exponent satisfies

$$\begin{aligned}\bar{D}_{\varepsilon^-}(\boldsymbol{X}\|\hat{\boldsymbol{X}}) &\leq \limsup_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \leq \bar{D}_{\varepsilon}(\boldsymbol{X}\|\hat{\boldsymbol{X}}) \\ \underline{D}_{\varepsilon^-}(\boldsymbol{X}\|\hat{\boldsymbol{X}}) &\leq \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \leq \underline{D}_{\varepsilon}(\boldsymbol{X}\|\hat{\boldsymbol{X}})\end{aligned}$$

where $\beta_n^*(\varepsilon)$ represents the minimum type-II error probability subject to a fixed type-I error bound $\varepsilon \in [0,1)$.

The general formula for Neyman-Pearson type-II error exponent subject to an exponential test level is also proved in [4, Theorem 3]. We, herein provide an extension of this result and express it in terms of the ε-inf/sup-divergence rates.

Theorem 3.5 (Neyman-Pearson type-II error exponent for an exponential test level)

Fix $s \in (0,1)$ and $\varepsilon \in [0,1)$. It is possible to choose decision regions for a binary hypothesis testing problem with arbitrary datawords of blocklength n, (which are governed by either the null

hypothesis distribution $P_{\boldsymbol{X}}$ or the alternative hypothesis distribution $P_{\hat{\boldsymbol{X}}}$), such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n \;\geq\; \bar{D}_{\varepsilon}(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}}) \quad \text{and} \quad \limsup_{n\to\infty} -\frac{1}{n}\log\alpha_n \;\geq\; \underline{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \tag{3.9}$$

or

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n \;\geq\; \underline{D}_{\varepsilon}(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}}) \quad \text{and} \quad \limsup_{n\to\infty} -\frac{1}{n}\log\alpha_n \;\geq\; \bar{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \tag{3.10}$$

where $\hat{\boldsymbol{X}}^{(s)}$ exhibits the tilted distributions $\{P^{(s)}_{\hat{X}^n}\}_{n=1}^{\infty}$ defined by

$$dP^{(s)}_{\hat{X}^n}(x^n) \triangleq \frac{1}{\Omega_n(s)}\exp\left\{s\log\frac{dP_{X^n}}{dP_{\hat{X}^n}}(x^n)\right\}dP_{\hat{X}^n}(x^n),$$

and

$$\Omega_n(s) \triangleq \int_{\mathcal{X}^n}\exp\left\{s\log\frac{dP_{X^n}}{dP_{\hat{X}^n}}(x^n)\right\}dP_{\hat{X}^n}(x^n).$$

Here, α_n and β_n are the type-I and type-II error probabilities respectively.

***Proof*:** For ease of notation, we use $\tilde{\boldsymbol{X}}$ to represent $\hat{\boldsymbol{X}}^{(s)}$. We only prove (3.9); (3.10) can be similarly demonstrated.

By definition of $dP^{(s)}_{\hat{X}^n}(\cdot)$, we have

$$\frac{1}{s}\left[\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|\hat{X}^n)\right] + \frac{1}{1-s}\left[\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|X^n)\right] = -\frac{1}{s(1-s)}\left[\frac{1}{n}\log\Omega_n(s)\right]. \tag{3.11}$$

Let $\bar{\Omega}\triangleq\limsup_{n\to\infty}(1/n)\log\Omega_n(s)$. Then, for any $\gamma>0$, $\exists N_0$ such that $\forall\ n>N_0$,

$$(1/n)\log\Omega_n(s) < \bar{\Omega}+\gamma.$$

From (3.11),

$$\begin{aligned}
\underline{d}_{\widetilde{X}^n\|\hat{X}^n}(\theta) &\triangleq \liminf_{n\to\infty} Pr\left\{\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|\hat{X}^n)\leq\theta\right\}\\
&= \liminf_{n\to\infty} Pr\left\{-\frac{1}{1-s}\left[\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|X^n)\right]-\frac{1}{s(1-s)}\left[\frac{1}{n}\log\Omega_n(s)\right]\leq\frac{\theta}{s}\right\}\\
&= \liminf_{n\to\infty} Pr\left\{\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|X^n)\geq-\frac{1-s}{s}\theta-\frac{1}{s}\left[\frac{1}{n}\log\Omega_n(s)\right]\right\}\\
&\leq \liminf_{n\to\infty} Pr\left\{\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|X^n)>-\frac{1-s}{s}\theta-\frac{1}{s}\bar{\Omega}-\frac{\gamma}{s}\right\}\\
&= 1-\limsup_{n\to\infty} Pr\left\{\frac{1}{n}d_{\widetilde{X}^n}(\widetilde{X}^n\|X^n)\leq-\frac{1-s}{s}\theta-\frac{1}{s}\bar{\Omega}-\frac{\gamma}{s}\right\}\\
&= 1-\bar{d}_{\widetilde{X}^n\|X^n}\left(-\frac{1-s}{s}\theta-\frac{1}{s}\bar{\Omega}-\frac{\gamma}{s}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\bar{D}_\varepsilon(\tilde{\boldsymbol{X}}\|\hat{\boldsymbol{X}}) &\stackrel{\triangle}{=} \sup\{\theta : \underline{d}_{\widetilde{X}^n\|\hat{X}^n}(\theta) \le \varepsilon\} \\
&\ge \sup\left\{\theta : 1 - \bar{d}_{\widetilde{X}^n\|\hat{X}^n}\left(-\frac{1-s}{s}\theta - \frac{1}{s}(\bar{\Omega}+\gamma)\right) < \varepsilon\right\} \\
&= \sup\left\{-\frac{1}{1-s}(\bar{\Omega}+\gamma) - \frac{s}{1-s}\theta' : \bar{d}_{\widetilde{X}^n\|X^n}(\theta') > 1-\varepsilon\right\} \\
&= -\frac{1}{1-s}(\bar{\Omega}+\gamma) - \frac{s}{1-s}\inf\{\theta' : \bar{d}_{\widetilde{X}^n\|X^n}(\theta') > 1-\varepsilon\} \\
&= -\frac{1}{1-s}(\bar{\Omega}+\gamma) - \frac{s}{1-s}\sup\{\theta' : \bar{d}_{\widetilde{X}^n\|X^n}(\theta') \le 1-\varepsilon\} \\
&= -\frac{1}{1-s}(\bar{\Omega}+\gamma) - \frac{s}{1-s}\underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X}).
\end{aligned}$$

Finally, choose the acceptance region for null hypothesis as

$$\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{\hat{X}^n}}(X^n) \ge \bar{D}_\varepsilon(\tilde{\boldsymbol{X}}\|\hat{\boldsymbol{X}})\right\}.$$

Therefore

$$\beta_n = P_{\hat{X}^n}\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{\hat{X}^n}}(X^n) \ge \bar{D}_\varepsilon(\tilde{\boldsymbol{X}}\|\hat{\boldsymbol{X}})\right\} \le \exp\left\{-n\bar{D}_\varepsilon(\tilde{\boldsymbol{X}}\|\hat{\boldsymbol{X}})\right\},$$

and

$$\begin{aligned}
\alpha_n &= P_{X^n}\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{\hat{X}^n}}(X^n) < \bar{D}_\varepsilon(\tilde{\boldsymbol{X}}\|\hat{\boldsymbol{X}})\right\} \\
&\le P_{X^n}\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{\hat{X}^n}}(X^n) < -\frac{1}{1-s}(\bar{\Omega}+\gamma) - \frac{s}{1-s}\underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X})\right\} \\
&= P_{X^n}\left\{\frac{-1}{1-s}\left[\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{X^n}}(X^n)\right] - \frac{1}{s(1-s)}\left[\frac{1}{n}\log\Omega_n(s)\right]\right. \\
&\qquad\qquad \left. < -\frac{\bar{\Omega}+\gamma}{s(1-s)} - \frac{1}{1-s}\underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X})\right\} \\
&= P_{X^n}\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{X^n}}(X^n) > \underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X}) + \frac{1}{s}\left[\bar{\Omega} - \frac{1}{n}\log\Omega_n(s)\right] + \frac{\gamma}{s}\right\}.
\end{aligned}$$

Then, for $n > N_0$,

$$\begin{aligned}
\alpha_n &\le P_{X^n}\left\{\frac{1}{n}\log\frac{dP_{\widetilde{X}^n}}{dP_{X^n}}(X^n) > \underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X})\right\} \\
&\le \exp\left\{-n\underline{D}_{1-\varepsilon}(\tilde{\boldsymbol{X}}\|\boldsymbol{X})\right\}.
\end{aligned}$$

Consequently,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n \ge \bar{D}_\varepsilon(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}}) \quad \text{and} \quad \limsup_{n\to\infty} -\frac{1}{n}\log\alpha_n \ge \underline{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}).$$

□

IV. Conclusions

In light of the new information quantiles introduced in [5], a generalized version of the Asymptotic Equipartition Property (AEP) is proved. General data compaction and compression (source coding) theorems for block codes and general expressions for the Neyman-Pearson hypothesis testing type-II error exponent are also derived.

Finally, it is demonstrated that by using these new quantities, Shannon's coding theorems can be reformulated in their *most general form* and the error probability of an *arbitrary* stochastic communication system can be determined.

Appendix A

Claim (*cf Proof of Theorem 2.3*)

$$\limsup_{n\to\infty} E_{\tilde{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n^*))] < \varepsilon.$$

***Proof*:**

step 1: Define

$$A_{n,\gamma}^{(\varepsilon)} \stackrel{\triangle}{=} \left\{(x^n, y^n) : \frac{1}{n}\rho_n(x^n, y^n) \le \bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X}, \tilde{\boldsymbol{Y}}) + \gamma,\ \frac{1}{n} i_{X^n Y^n}(x^n, y^n) \le \bar{I}(\boldsymbol{X}; \tilde{\boldsymbol{Y}}) + \gamma\right\}.$$

Since

$$\liminf_{n\to\infty} Pr\left(D \stackrel{\triangle}{=} \left\{\frac{1}{n}\rho_n(X^n, \tilde{Y}^n) \le \bar{\Lambda}_{1-\varepsilon}(\boldsymbol{X}, \tilde{\boldsymbol{Y}}) + \gamma\right\}\right) > 1 - \varepsilon,$$

and

$$\liminf_{n\to\infty} Pr\left(E \stackrel{\triangle}{=} \left\{\frac{1}{n} i_{X^n \tilde{Y}^n}(X^n; \tilde{Y}^n) \le \bar{I}(\boldsymbol{X}; \tilde{\boldsymbol{Y}}) + \gamma\right\}\right) = 1,$$

we have

$$\begin{aligned}
\liminf_{n\to\infty} Pr(A_{n,\gamma}^{(\varepsilon)}) &= \liminf_{n\to\infty} Pr(D \cap E) \\
&\ge \liminf_{n\to\infty} Pr(D) + \liminf_{n\to\infty} Pr(E) - 1 \\
&> (1-\varepsilon) + 1 - 1 = 1 - \varepsilon.
\end{aligned}$$

step 2: Let $K(x^n, y^n)$ be the indicator function of $A_{n,\gamma}^{(\varepsilon)}$:

$$K(x^n, y^n) = \begin{cases} 1, & \text{if } (x^n, y^n) \in A_{n,\gamma}^{(\varepsilon)}; \\ 0, & \text{otherwise.} \end{cases}$$

step 3: By following a similar argument in [9, equations (9)-(12)], we obtain,

$$\begin{aligned}
& E_{\tilde{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n^*))] \\
= & \sum_{\mathcal{C}_n^*} P_{\tilde{Y}^n}(\mathcal{C}_n^*) \sum_{x^n \notin A(\mathcal{C}_n^*)} P_{X^n}(x^n) \\
= & \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \sum_{\mathcal{C}_n^* : x^n \notin A(\mathcal{C}_n^*)} P_{\tilde{Y}^n}(\mathcal{C}_n^*)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \left(1 - \sum_{y^n \in \mathcal{Y}^n} P_{\tilde{Y}^n}(y^n) K(x^n, y^n)\right)^M \\
&\leq \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \left(1 - e^{-n(\bar{I}(\boldsymbol{X};\tilde{\boldsymbol{Y}})+\gamma)} \times \sum_{y^n \in \mathcal{Y}^n} P_{\tilde{Y}^n|X^n}(y^n|x^n) K(x^n, y^n)\right)^M \\
&\leq 1 - \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} P_{X^n}(x^n) P_{\tilde{Y}^n|X^n}(x^n, y^n) K(x^n, y^n) + \exp\left\{-e^{n(R-R_{1-\varepsilon}(D)-\gamma)}\right\}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\limsup_{n\to\infty} E_{\tilde{Y}^n}[P_{X^n}(A^n(\mathcal{C}_n^*))] &\leq 1 - \liminf_{n\to\infty} Pr(A_{n,\gamma}^{(\varepsilon)}) \\
&< 1 - (1-\varepsilon) = \varepsilon.
\end{aligned}
$$

□

Appendix B

Claim: Fix $\varepsilon \in [0,1)$. $T_\varepsilon(\boldsymbol{X})$ is right-continuous in ε.

***Proof*:**

Suppose $T_\varepsilon(\boldsymbol{X})$ is not right-continuous for some $\varepsilon \in [0,1)$. Then there exists $\gamma > 0$ such that

$$
T_{\varepsilon+\delta}(\boldsymbol{X}) < T_\varepsilon(\boldsymbol{X}) + 3\gamma \text{ for every } 1-\varepsilon > \delta > 0,
$$

which guarantees the existence of R satisfying

$$
T_{\varepsilon+\delta}(\boldsymbol{X}) < R - \gamma < R < T_\varepsilon(\boldsymbol{X})
$$

for every $1-\varepsilon > \delta > 0$. Hence, $R-\gamma$ is $(\varepsilon+\delta)$-achievable for every $1-\varepsilon > \delta > 0$, and R is not ε-achievable.

By definition of $(\varepsilon+\delta)$-achievability, there exists a code $D_n(\delta)$ such that

$$
\limsup_{n\to\infty} (1/n) \log |D_n(\delta)| = R - \gamma \text{ and } \limsup_{n\to\infty} Pe(D_n(\delta)) \leq \varepsilon + \delta.
$$

Therefore, there exists $M(\delta)$ such that for $n > M(\delta)$,

$$
(1/n) \log |D_n(\delta)| < R \text{ and } Pe(D_n(\delta)) < \varepsilon + 2\delta.
$$

Observe that if we increase the code size of $D_n(\delta)$ to obtain a new code $D'_n(\delta)$ with $(1/n)\log|D'_n(\delta)| = R$ for $n > M(\delta)$, then the error probability will not increase, i.e.,

$$P_e(D'_n(\delta)) < \varepsilon + 2\delta.$$

Now define a new code E_n as follows:

$$\begin{aligned}
E_n &= D'_n(\delta) \text{ for } M(\delta) < n \leq \max\{M(\delta), M(\delta/2)\} \\
E_n &= D'_n(\delta/2) \text{ for } \max\{M(\delta), M(\delta/2)\} < n \leq \max\{M(\delta), M(\delta/2), M(\delta/3)\} \\
E_n &= D'_n(\delta/3) \text{ for } \max\{M(\delta), M(\delta/2), M(\delta/3)\} < n \leq \max\{M(\delta), M(\delta/2), M(\delta/3), M(\delta/4)\} \\
&\vdots
\end{aligned}$$

Then for $n > M(\delta)$, $(1/n)\log|E_n| = R$ but $\limsup_{n\to\infty} Pe(E_n) \leq \varepsilon$, contradicting the fact that R is not ε-achievable. □

Claim: $T_{1-\varepsilon}(\boldsymbol{X}) = \bar{H}_{\varepsilon^-}(\boldsymbol{X})$ for $\varepsilon \in (0,1]$ and $T_1(\boldsymbol{X}) = 0$.

***Proof*:** The first result is an immediate consequence of the right-continuity of $T_{1-\varepsilon}(\boldsymbol{X})$ w.r.t. $(1-\varepsilon) \in [0,1)$. $T_1(\boldsymbol{X})$, by definition, is the infimum of the 1-achievable data compaction rate which requires the existence of codes $\mathscr{C}_n$ with

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathscr{C}_n| = R,$$

and

$$\limsup_{n\to\infty} Pe(\mathscr{C}_n) \leq 1.$$

We can then choose an empty code set, and obtain $T_1(\boldsymbol{X}) = 0$. □

Acknowledgment

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

References

1. F. Alajaji and T. Fuja. "A communication channel modeled on contagion," *IEEE Trans. Info. Theory*, IT–40(6):2035–2041, November (1994).

2. R. E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, Massachusetts (1988).

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

4. P.-N. Chen, "General formulas for the Neyman-Pearson type-II error exponent subject to fixed and exponential type-I error bounds," *IEEE Trans. Info. Theory*, IT–42(1):316–323, January (1996).

5. P.-N. Chen and F. Alajaji, "Generalized source coding theorems and hypothesis testing: Part I – Information measures," *Journal of the Chinese Institute of Engineers,* to appear, May (1998).

6. I. Csiszár, "Information theory and ergodic theory," *Problems of Control and Inform. Theory*, 16(1):3–27 (1987).

7. R. M. Gray, *Entropy and Information Theory.* Springer-Verlag, New York (1990).

8. T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Info. Theory*, IT–39(3):752–772, May (1993).

9. Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Info. Theory*, IT–42(1):63–86, Jan. (1996).

10. S. Vembu, S. Verdú and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Info. Theory*, IT–41(1):44–54, Jan. (1995).

11. S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Info. Theory*, IT–40(4):1147–1157, Jul. (1994).

Nomenclature

$(1-\varepsilon)$-achievable data compaction rate	$T_{1-\varepsilon}(\boldsymbol{X})$
$(1-\varepsilon)$-achievable data compression rate at distortion D	$T_{1-\varepsilon}(D, \boldsymbol{X})$
δ-inf-divergence rate	$\underline{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}})$
δ-inf-entropy rate	$\underline{H}_\delta(\boldsymbol{X})$
δ-inf-information rate	$\underline{I}_\delta(\boldsymbol{X};\boldsymbol{Y})$
δ-sup-divergence rate	$\bar{D}_\delta(\boldsymbol{X}\|\hat{\boldsymbol{X}})$
δ-sup-entropy rate	$\bar{H}_\delta(\boldsymbol{X})$
δ-sup-information rate	$\bar{I}_\delta(\boldsymbol{X};\boldsymbol{Y})$
ε-sup-distortion rate	$\bar{\Lambda}_\varepsilon(\boldsymbol{X},\boldsymbol{Y})$
ϵ-capacity	C_ϵ
channel capacity	C
channel transition distribution	$P_{W^n} = P_{Y^n\|X^n}$
distortion inf-spectrum	$\underline{\lambda}_{(\boldsymbol{X},f(\boldsymbol{X}))}(\theta)$
divergence inf-spectrum	$\underline{d}_{\boldsymbol{X}\|\hat{\boldsymbol{X}}}(\theta)$
divergence sup-spectrum	$\bar{d}_{\boldsymbol{X}\|\hat{\boldsymbol{X}}}(\theta)$
entropy density	$h_{X^n}(X^n)$
entropy inf-Spectrum	$\underline{h}_{\boldsymbol{X}}(\theta)$
entropy sup-Spectrum	$\bar{h}_{\boldsymbol{X}}(\theta)$
inf-divergence rate	$\underline{D}(\boldsymbol{X}\|\hat{\boldsymbol{X}})$
inf-entropy rate	$\underline{H}(\boldsymbol{X})$
inf-information rate	$\underline{I}(\boldsymbol{X};\boldsymbol{Y})$
information density	$i_{X^nW^n}(x^n;y^n)$
information inf-spectrum	$\underline{i}_{(\boldsymbol{X},\boldsymbol{Y})}(\theta)$
information sup-spectrum	$\bar{i}_{(\boldsymbol{X},\boldsymbol{Y})}(\theta)$
input alphabet	$\mathcal{A}$
input distributions	P_{X^n}
log-likelihood ratio	$d_{X^n}(X^n\|\hat{X}^n)$
output alphabet	$\mathcal{B}$
sup-divergence rate	$\bar{D}(\boldsymbol{X}\|\hat{\boldsymbol{X}})$
sup-entropy rate	$\bar{H}(\boldsymbol{X})$
sup-information rate	$\bar{I}(\boldsymbol{X};\boldsymbol{Y})$