A Rate-Distortion Theorem for Arbitrary Discrete Sources

Po-Ning Chen and Fady Alajaji

***Abstract*— A rate-distortion theorem for arbitrary (not necessarily stationary or ergodic) discrete-time finite-alphabet sources is given. This result, which provides the expression of the minimum ϵ-achievable fixed-length coding rate subject to a fidelity criterion, extends a recent data compression theorem by Steinberg and Verdú.**

***Index Terms*—Arbitrary discrete sources, data compression, rate-distortion theory, Shannon theory.**

I. Introduction

We consider the problem of source coding with a fidelity criterion for arbitrary (not necessarily stationary or ergodic) discrete-time finite-alphabet sources. We prove a general rate-distortion theorem by establishing the expression of the minimum ϵ-achievable block coding rate subject to a fidelity criterion.

In [3, Theorem 10, part a)], Steinberg and Verdú demonstrate a data compression theorem for arbitrary sources under the restriction that the probability of excessive distortion due to the achievable data compression codes is asymptotically equal to zero (cf. [3, Definitions 30 and 31]). In this work, we provide a variant of their result by relaxing the restriction on the probability of excessive distortion (cf. (3.1)).

II. Preliminaries

Consider a random process $\boldsymbol{X}$ defined by a sequence of finite-dimensional distributions [2]

$$\boldsymbol{X} \triangleq \{X^n = (X_1^{(n)}, \cdots, X_n^{(n)})\}_{n=1}^{\infty}.$$

Let

$$\boldsymbol{Y} \triangleq \{Y^n = (Y_1^{(n)}, \cdots, Y_n^{(n)})\}_{n=1}^{\infty}$$

be the corresponding output process induced by $\boldsymbol{X}$ via the channel

$$\boldsymbol{W} \triangleq \{W_{Y^n|X^n} = P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty},$$

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. We assume that $\mathcal{X}$ to $\mathcal{Y}$ are finite.

Definition 2.1 ([2]): Given a joint distribution $P_{X^nY^n} = W_{Y^n|X^n}P_{X^n}$ on $\mathcal{X}^n \times \mathcal{Y}^n$ with marginals P_{X^n} and P_{Y^n}, the *information density* is defined by

$$i_{X^nY^n}(a^n; b^n) = \log \frac{W_{Y^n|X^n}(b^n \mid a^n)}{P_{Y^n}(b^n)}.$$

Manuscript received April 15, 1996; revised January 20, 1998. The work of P.-N. Chen was supported in part by National Chiao-Tung University. The work of F. Alajaji was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant OGP0183645. The material in this correspondence was presented in part at the International Symposium on Information Theory and Its Applications, Victoria, BC, Canada, September 1996.

P.-N. Chen is with the Department of Communication Engineering, National Chiao-Tung University, HsinChu, Taiwan, R.O.C.

F. Alajaji is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ont. K7L 3N6, Canada.

Publisher Item Identifier S 0018-9448(98)03468-3.

Definition 2.2 ([2], [3]): The *sup-information rate* $\bar{I}(\boldsymbol{X};\boldsymbol{Y})$ of the joint process $\boldsymbol{XY}$ is defined as the *limsup in probability*[1] of the sequence of normalized information densities $\frac{1}{n}\, i_{X^nY^n}(X^n;Y^n)$.

Analogously, the *inf-information rate* $\underline{I}(\boldsymbol{X};\boldsymbol{Y})$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as the *liminf in probability* of the sequence of normalized information densities $\frac{1}{n}\, i_{X^nY^n}(X^n;Y^n)$.

When $\boldsymbol{X}$ is equal to $\boldsymbol{Y}$, $\bar{I}(\boldsymbol{X};\boldsymbol{X})$ (respectively, $\underline{I}(\boldsymbol{X};\boldsymbol{X})$) is referred to as the *sup* (respectively, *inf*) *entropy rate* of $\boldsymbol{X}$ and is denoted by $\bar{H}(\boldsymbol{X})$ (respectively, $\underline{H}(\boldsymbol{X})$).

Definition 2.3 ([2], [3]): Given a joint distribution $P_{X^nY^n} = W_{Y^n|X^n}P_{X^n}$, the *conditional entropy density* is defined by

$$i_{Y^n|X^n}(b^n \mid a^n) = -\log W_{Y^n|X^n}(b^n \mid a^n).$$

The *conditional sup-entropy rate* $\bar{H}(\boldsymbol{Y} \mid \boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is defined as the limsup in probability of the sequence of normalized conditional entropy densities $\frac{1}{n}\, i_{Y^n|X^n}(Y^n \mid X^n)$.

Analogously, the *conditional inf-entropy rate* $\underline{H}(\boldsymbol{Y} \mid \boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is defined as the liminf in probability of $\frac{1}{n}\, i_{Y^n|X^n}(Y^n \mid X^n)$.

III. General Data Compression Theorem

Definition 3.1 (e.g., [1]): Given a finite source alphabet $\mathcal{X}$ and a finite 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 reproducing letters.

Definition 3.2: A distortion measure $\rho_n(\cdot,\cdot)$ is a mapping

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

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

Definition 3.3: Let $\boldsymbol{X}$ and $\{\rho_n(\cdot,\cdot)\}_{n\geq 1}$ be given. Let

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

denote a sequence of data compression codes for $\boldsymbol{X}$. The *distortion spectrum* $\underline{\lambda}_{\boldsymbol{X}\boldsymbol{f}(\boldsymbol{X})}(\theta)$ for $\boldsymbol{f}(\cdot)$ is defined by

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

Definition 3.4: Fix $D > 0$ and $1 > \varepsilon > 0$. R is an *ε-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

$$\sup[\theta : \underline{\lambda}_{\boldsymbol{X}\boldsymbol{f}(\boldsymbol{X})}(\theta) \leq \varepsilon] \leq D. \tag{3.1}$$

[1] If A_n is a sequence of random variables, then its *liminf in probability* is the largest extended real number α such that for all $\xi > 0$,

$$\lim_{n\to\infty} \Pr[A_n \leq \alpha - \xi] = 0.$$

Similarly, its *limsup in probability* is the smallest extended real number β such that for all $\xi > 0$ [2]

$$\lim_{n\to\infty} \Pr[A_n \geq \beta + \xi] = 0.$$

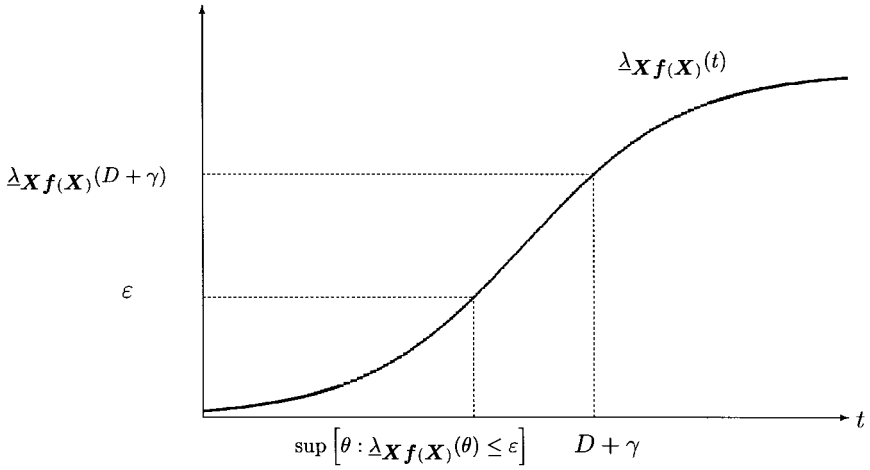

Fig. 1. $\underline{\lambda}_{\boldsymbol{X}f(\boldsymbol{X})}(D+\gamma) > \varepsilon \Rightarrow \sup[\theta : \underline{\lambda}_{\boldsymbol{X}f(\boldsymbol{X})}(\theta) \le \varepsilon] \le D+\gamma$.

Note that (3.1) is equivalent to stating that the limsup of the probability of excessive distortion (i.e., distortion larger than D) is smaller than $1-\varepsilon$.

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

Theorem 3.1 (General Data Compression Theorem): Fix $D > 0$ and $1 > \varepsilon > 0$. Let $\boldsymbol{X}$ and $\{\rho_n(\cdot,\cdot)\}_{n\ge 1}$ be given. Then

$$T_\varepsilon(D, \boldsymbol{X}) = R_\varepsilon(D)$$

where

$$R_\varepsilon(D) \triangleq \inf_{\{\boldsymbol{W}:\sup[\theta:\underline{\lambda}_{\boldsymbol{XY}}(\theta)\le\varepsilon]\le D\}} \bar{I}(\boldsymbol{X};\boldsymbol{Y})$$

where the infimum is taken over all conditional distributions $\boldsymbol{W} = \{P_{Y^n|X^n}\}_{n=1}^\infty$ for which the joint distribution $P_{\boldsymbol{XY}} = P_{\boldsymbol{X}}\boldsymbol{W}$ satisfies the distortion constraint.

Proof:

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

$$\limsup_{n\to\infty} \frac{1}{n}\|f_n\| \le R_\varepsilon(D) + 2\gamma$$

and

$$\sup[\theta : \underline{\lambda}_{\boldsymbol{X}f(\boldsymbol{X})}(\theta) \le \varepsilon] \le D+\gamma.$$

Step 1: Let $\tilde{\boldsymbol{W}}$ be the channel distribution achieving $R_\varepsilon(D)$, and let $P_{\tilde{\boldsymbol{Y}}}$ be the $\boldsymbol{Y}$-marginal of $P_{\boldsymbol{X}}\tilde{\boldsymbol{W}}$.

Step 2: Let $R = R_\varepsilon(D) + 2\gamma$. Choose $M = e^{nR}$ n-blocks independently according to $P_{\tilde{\boldsymbol{Y}}}$, 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

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

Step 4: *Claim:*

$$\limsup_{n\to\infty} E_{\tilde{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n))] < 1-\varepsilon.$$

The proof of this claim is provided in the Appendix. Therefore, there exists (a sequence of) $\mathcal{C}_n^*$ such that

$$\limsup_{n\to\infty} P_{X^n}(A^c(\mathcal{C}_n^*)) < 1-\varepsilon.$$

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

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

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,

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

Hence

$$\sup[\theta : \underline{\lambda}_{\boldsymbol{X}f(\boldsymbol{X})}(\theta) \le \varepsilon] \le D+\gamma$$

where the last step is clearly depicted in Fig. 1.

This proves the forward part.

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

$$\sup[\theta : \underline{\lambda}_{\boldsymbol{X}f(\boldsymbol{X})}(\theta) \le \varepsilon] \le D$$

then

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

Let

$$\hat{W}^n(y^n \mid x^n) \triangleq \begin{cases} 1, & \text{if } y^n = f_n(x^n) \\ 0, & \text{otherwise.} \end{cases}$$

Let $\hat{Y}^n$ denote the output corresponding to input X^n and channel $\hat{W}^n$. Then to evaluate the statistical properties of the random sequence $\{(1/n)\rho_n(X^n, f_n(X^n)\}_{n=1}^\infty$ under distribution

P_{X^n} is equivalent to evaluating those of the random sequence $\{(1/n)\rho_n(X^n, \hat{Y}^n)\}_{n=1}^{\infty}$ under distribution $P_{X^n}\hat{W}^n$. Therefore,

$$\begin{aligned} R_\varepsilon(D) &\triangleq \inf_{\{\boldsymbol{W}:\sup[\theta:\underline{\lambda}_{\boldsymbol{XY}}(\theta)\le\varepsilon]\le D\}} \bar{I}(\boldsymbol{X};\boldsymbol{Y}) \\ &\le \bar{I}(\boldsymbol{X};\hat{\boldsymbol{Y}}) \\ &\le \bar{H}(\hat{\boldsymbol{Y}}) - \underline{H}(\hat{\boldsymbol{Y}} \mid \boldsymbol{X}) \\ &\le \bar{H}(\hat{\boldsymbol{Y}}) \\ &\le \limsup_{n\to\infty} \frac{1}{n}\log\|f_n\|, \end{aligned}$$

where the second inequality follows from [4, Theorem 8, property (d)] and the third inequality follows from the fact that $\underline{H}(\hat{\boldsymbol{Y}} \mid \boldsymbol{X}) \ge 0$. □

APPENDIX

Claim (cf. Proof of Theorem 3.1):

$$\limsup_{n\to\infty} E_{\tilde{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n^*))] < 1-\varepsilon.$$

Proof:

Step 1: Let

$$D^{(\varepsilon)} \triangleq \sup\{\theta : \underline{\lambda}_{\boldsymbol{X}\tilde{\boldsymbol{Y}}}(\theta) \le \varepsilon\}.$$

Define

$$A_{n,\gamma}^{(\varepsilon)} \triangleq \left\{(x^n, y^n) : \frac{1}{n}\rho_n(x^n, y^n) \le D^{(\varepsilon)} + \gamma \text{ and } \frac{1}{n} i_{X^nY^n}(x^n, y^n) \le \bar{I}(\boldsymbol{X};\tilde{\boldsymbol{Y}}) + \gamma\right\}.$$

Since

$$\liminf_{n\to\infty} \Pr\left(\mathcal{D} \triangleq \left\{\frac{1}{n}\rho_n(X^n, \tilde{Y}^n) \le D^{(\varepsilon)} + \gamma\right\}\right) > \varepsilon$$

and

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

we have

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

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 [3, Eqs. (9)–(12)], we obtain

$$\begin{aligned} &E_{\tilde{\boldsymbol{Y}}}[P_{X^n}(A^c(\mathcal{C}_n^*))] \\ &= \sum_{\mathcal{C}_n^*} P_{\tilde{Y}^n}(\mathcal{C}_n^*) \sum_{x^n \notin A(\mathcal{C}_n^*)} P_{X^n}(x^n) \\ &= \sum_{x^n\in\mathcal{X}^n} P_{X^n}(x^n) \sum_{\mathcal{C}_n^*: x^n \notin A(\mathcal{C}_n^*)} P_{\tilde{Y}^n}(\mathcal{C}_n^*) \\ &= \sum_{x^n\in\mathcal{X}^n} P_{X^n}(x^n)\left(1 - \sum_{y^n\in\mathcal{Y}^n} P_{\tilde{Y}^n}(y^n)K(x^n, y^n)\right)^M \\ &\le \sum_{x^n\in\mathcal{X}^n} P_{X^n}(x^n)\Bigg(1 - e^{-n(\bar{I}(\boldsymbol{X};\tilde{\boldsymbol{Y}})+\gamma)} \\ &\quad\times \sum_{y^n\in\mathcal{Y}^n} P_{\tilde{Y}^n|X^n}(y^n \mid x^n)K(x^n, y^n)\Bigg)^M \\ &\le 1 - \sum_{x^n\in\mathcal{X}^n}\sum_{y^n\in\mathcal{Y}^n} P_{X^n}(x^n)P_{\tilde{Y}^n|X^n}(x^n, y^n)K(x^n, y^n) \\ &\quad + \exp\{-e^{n(R-R_\varepsilon(D)-\gamma)}\}. \end{aligned}$$

Therefore,

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

□

ACKNOWLEDGMENT

The authors would like to thank Prof. S. Verdú for his valuable advice and encouragements.

REFERENCES

[1] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison Wesley, 1988.
[2] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.
[3] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. I42, pp. 63–86, Jan. 1996.
[4] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.

On One Useful Inequality in Testing of Hypotheses

Marat V. Burnashev

***Abstract*—A simple proof of one probabilistic inequality is presented.**

***Index Terms*—Error probabilities, testing of hypotheses.**

I. MAIN INEQUALITY

Let P and Q be two given probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. We consider testing of hypotheses P and Q using one observation. For an arbitrary decision rule, let α and β denote the two kinds of error probabilities. If both error probabilities have equal costs (or we want to minimize the maximum of them) then it is natural to investigate the minimal possible sum $\inf\{\alpha+\beta\}$ for the best decision rule.

Manuscript received July 10, 1997; revised November 24, 1997. This work was supported by the Russian Foundation for Fundamental Research under Grants N 95-01-00136a and INTAS-94-469.

The author is with the Institute for Problems of Information Transmission, Russian Academy of Sciences, 101447 Moscow, Russia.

Publisher Item Identifier S 0018-9448(98)03470-1.