

**Compression of Information Sources with
Memory:
A Tutorial Review**

By
Sundeep Venkatraman
Department of Electrical Engineering
University of Notre Dame

Introduction:

In the early days of the development of information theory, sources were often conveniently modeled as memoryless processes and the analysis was based on the law of large numbers. However, it was soon apparent that such a model was not realistic enough and hence, a more general model of the information source as a stationary random process was adopted, where the ergodic theorem was used instead of the law of large numbers.

The use of the ergodic theorem to prove the entropy rate theorem, better known as the Asymptotic Equipartition Property was one of the key results obtained from this new point of view ([1]). The importance of ergodic theory however, extends beyond the use of the ergodic theorem for proving the AEP. Certain results from ergodic theory can be applied to the theory of Universal source coding as well ([2],[3]). Some of these results are presented in the sections to come.

While the ergodic theorem is a powerful tool in the attempt to extend the basic results of information theory pertaining to source coding, there have also other attempts made to obtain general results for the case of nonstationary sources ([4],[5]). Some of the pertinent results from these references are also presented in later sections.

Application of the ergodic theorem to the analysis of source processes.

The following is the definition of the property of Ergodicity [6]

Let $f(x) \in L^1_\mu(m)$, where m is a measurable space and let μ be the finite measure over which f is integrable. If T refers to any measurable transformation $T: m \rightarrow m$ Then T is said to be ergodic if the following relation holds true with probability 1.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \hat{f}(x) = \int_m f(x) d\mu(x) \quad \text{almost everywhere} \quad (1)$$

(Note: Often, T^k is a left shift by k .)

In the theory of probability, similar assertions are referred to as the law of large numbers and since the convergence takes place almost everywhere, the ergodic theorem is actually a stronger result than the law of large numbers.

One of the important results obtained by the use of the ergodic theorem in place of the law of large numbers is in [1], where a proof is given for the Asymptotic Equipartition Property (AEP) for ergodic finite alphabet sources in the convergence in probability sense. The statement of the property as given in this paper is "Every ergodic source has the AEP". According to the proof,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log(\mu[0, n-1; X]) \rightarrow \mathbf{H}(X) \quad (\text{in probability})$$

Where $\mathbf{H}(X)$ is the entropy rate of the source (since $\mu[0, n-1; X]$ represents the vector of random variables in X). In the special case where X_1, X_2, \dots, X_n are i.i.d random variables, the entropy rate equals the entropy and hence, we get the well known standard result for the i.i.d case.

For ergodic sources, since the entropy rate is both finite and a constant, we can replace the AEP used in the proofs (such as those for the fixed to fixed length source coding and the channel coding theorem) for an IID source process with the above result and hence prove the same for ergodic sources.

A result similar to the one in [1] is given in [2]. However, the approach used in [2] is novel in that it is presented more from an ergodic theory point of view than an information theory point of view and makes use of a few lemmas from ergodic theory to prove the AEP in the almost sure sense of convergence. Therefore, it constitutes a proof of strong typicality for ergodic sources which is a very important result. This paper also reviews some other results and ideas from ergodic theory which are of use in the analysis of universal source coding algorithms. Besides this, [2] also gives the following new interpretation of the typical set based on the following definition of a built-up set:

Def: A sequence x_1^N is said to be $(1-\delta)$ built-up from a set B if it can be expressed as a concatenation of variable length blocks

i.e. $x_1^N = b_1 b_2 \dots b_k$ where the total length of the b_i that belong to B is at least $(1-\delta)N$.

The precise formulation of the new typicality idea is as follows. The AEP provides an l and a set $C \subset A^l$ (where A is the source alphabet) such that $P(C) > 1-\delta/2$ and $|C| \leq 2^{l(H+\delta)}$ where δ is a free parameter. For each k , let S_k be the set of x_1^k that are $(1-\delta)$ built-up from C . If δ is small enough, then $|S_k| \leq 2^{k(H+\theta)}$ for all k ($\theta > 0$). The sets $\{S_k\}$ can be regarded as strongly typical in an empirical sense.

Ergodic decomposition of a Stationary Discrete Random Process.

Another interesting and useful result for the case of stationary but non ergodic sources called ergodic decomposition is found in [3] which states that a stationary source can be decomposed into a class of stationary ergodic sources. However, the ergodic source out of this class of sources which is in effect is itself a random variable (denoted by Θ). One of the possible uses of this result would be to apply it along with the result of [1] to non ergodic source processes to get a bound on rate of the source coder.

The ergodic decomposition can also be applied to obtain results for universal source codes. According to [3], an N -length code C_N , consists of a uniquely decipherable mapping of each source N -tuple x^N into a codeword of length $l_{C_N}(x^N)$. This mapping must be done without knowledge of θ , the index of the actual source in effect. The resulting average length of the code C , if the θ th source is in effect, is given by

$$l(C_N | \theta) = E_{\mu_\theta} \{l_{C_N}\} = \int_{\Omega} l_{C_N}(x^N(\omega)) d\mu_\theta(\omega)$$

It is also known that

$$l(C_N | \theta) \geq H(X^N | \theta)$$

The conditional redundancy of the code is defined as

$$r_N(C_N, \theta) = N^{-1}(l(C_N | \theta) - H(X^n | \theta)) \geq 0$$

The average redundancy of a code C_N is then given by

$$R_N(W, C_N) = \int_{\Lambda} r_N(C_N, \theta) dW(\theta)$$

If there exist a sequence of codes $\{C_N\}_{N=1}^{\infty}$ such that $\lim_{N \rightarrow \infty} R_N(W, C_N) = 0$, Then the sequence $\{C_N\}$ is said to be weighted universal. Likewise, A sequence of codes $\{C_N\}$ is said to be weakly minimax if for all θ , $\lim_{N \rightarrow \infty} r_N(W, C_N) = 0$

The following existence theorems for Noiseless Variable length codes are proved in [3].

1. There exist weighted universal codes for arbitrary classes of discrete alphabet discrete time stationary sources provided

$$I(X_n; \Theta | X^{n-1}) < \infty$$

2. There exists a sequence of weakly minimax universal codes if there exists a measure μ_o such that

$$E_{\nu_\theta}[-\ln(\mu_o(X_o))] < \infty \text{ for all } \nu_\theta \in M,$$

where M is a class of discrete time ergodic sources.

The following theorems for codes with a fidelity criterion are proved in [3]

1. Given a distortion function $\rho(C_N | \mu)$ and two stationary sources $[\Omega, \mu_\theta]$ and $[\Omega, \mu_\phi]$, for any integer N and codebook C_N , then

$$|\rho(C_N | \mu_\theta) - \rho(C_N | \mu_\phi)| \leq \rho_M 2^N d(\mu_\theta, \mu_\phi) \text{ (d - distribution distance)}$$

2. A sequence of weakly minimax universal source codes subject to a fidelity criterion exist for a class M of all discrete time ergodic sources having as alphabet a denumerable metric space and therefore by the ergodic decomposition, for the class of all discrete alphabet discrete time sources with such an alphabet.

General results for non-stationary source processes.

While the previously mentioned results were all valid for ergodic source processes or for stationary processes which can be decomposed into a class of stationary ergodic sources, attempts have also been made to analyze and obtain general expressions for optimum source coding rate for non-stationary sources. We shall review some of these attempts in this section.

Consider the following adaptive block to variable length coding scheme, given in [4]. If X_0, X_1, \dots refer to non-overlapping blocks of length N of the source sequence, each random block can be

encoded via a one-one mapping into a random codeword Y_i . The entire scheme is adaptive because the mapping from X_i to Y_i depends on the previously observed values of X_0, X_1, \dots, X_{i-1} .

The rate $r_X(\tau)$ at which the scheme $\tau \in C_N$ codes the given source X is defined by

$$r_X(\tau) = \limsup_{n \rightarrow \infty} n^{-1} \sum_{i=0}^{n-1} \left(\frac{EL(Y_i)}{N} \right)$$

The optimum rate is then given by

$$\inf_N \inf \{r_X(\tau) : \tau \in C_N\}$$

The stationary hull $\Lambda(X)$ of the source is the class of stationary processes $Z = (Z_0, Z_1, \dots)$ With alphabet A such that for some sequence of positive numbers $n_0 < n_1 < \dots$,

$$\lim_{j \rightarrow \infty} \frac{1}{n_j} \sum_{i=0}^{n_j-1} E(f(X_i, X_{i+1}, \dots)) = E[f(Z)]$$

Every process in the stationary hull is necessarily stationary (hence the name). The result proved in [4] for the optimal rate is

$$r_X = \sup \{H(Z) : Z \in \Lambda(X)\}, \quad \text{where } H(Z) \text{ is the entropy rate of } Z$$

This result can be justified considering the fact that the stationary hull is a kind of equivalent representation of the non-stationary process in terms of a class of stationary processes. In case X is itself stationary, $\Lambda(X)$ consists of X alone and the result is the same as the familiar result for stationary processes.

Another approach to the general case of non-stationary sources has been explored in [5]. The following definitions are pertinent.

Def 1. A channel W with input alphabet A and output alphabet B is a sequence of conditional distributions given by

$$\mathbf{W} = \{W^n(y^n | x^n) = P_{Y^n|X^n}(y^n | x^n); (x^n, y^n) \in A^n \times B^n\}_{n=1}^{\infty}$$

Def 2. Given a joint distribution $P_{X^n Y^n}(x^n, y^n)$, the information density is the function defined on $A^n \times B^n$:

$$i_{X^n W^n}(a^n, b^n) = \log \frac{W^n(b^n | a^n)}{P_{Y^n}(b^n)}$$

The distribution of $(1/n)i_{X^n W^n}(X^n, Y^n)$ is referred to as the information spectrum. The expected value of the information spectrum is the normalized mutual information $(1/nI(X^n; Y^n))$.

Def 3. The sup (resp. inf) information rate is defined as *limsup* (resp. *liminf*) in probability of the sequence of random variables $\{(1/n)i_{X^n W^n}(x^n, y^n)\}$ denoted by $\bar{I}(X; Y)$ ($\underline{I}(X; Y)$) (Hereafter, X and Y denote the vector of these random variables)

Also $\mathbf{I}(X; Y) = \lim_{n \rightarrow \infty} 1/nI(X^n; Y^n)$

Therefore for a channel with a finite alphabet, if $\bar{I}(X;Y) = \underline{I}(X;Y)$

We have $\mathbf{I}(X;Y) = \underline{I}(X;Y) = \bar{I}(X;Y)$

Def 4. For any positive integer M, a probability distribution P is said to be M-type if

$$P(\omega) \in \{0, 1/M, 2/M, \dots, 1\}$$

Def 5. The resolution R(P) of a probability distribution is the minimum log(M) such that P is M-type. Note: $H(P) \leq R(P)$

Def 6. Let $\varepsilon \geq 0$. R is an ε -achievable resolution rate for channel W if for every input process X and for all $\gamma > 0$ there exists an \tilde{X} whose resolution satisfies

$$\frac{1}{n} R(\tilde{X}^n) < R + \gamma$$

and $d(Y^n, \tilde{Y}^n) < \varepsilon$

The minimum ε -achievable resolution rate (resp. resolution rate) is called ε -resolvability (resp. resolvability) of the channel and is denoted by S_ε (resp. S). If the resolution rates are defined for a particular input process X, then we refer to them as $S_\varepsilon(X)$ (resp. S(X)). If we replaced resolution by entropy in the above definition, we get mean resolvability $\bar{S}_\varepsilon(X)$ (resp. $\bar{S}(X)$)

Def 7. R is an ε -achievable (fixed length) source coding rate for X if for all $\gamma > 0$ and sufficiently large n, there exists a collection of M n-tuples $\{x_1^n, \dots, x_M^n\}$ such that

$$1/n \log(M) < R + \gamma$$

and $P[X^n \notin \{x_1^n, \dots, x_M^n\}] \leq \varepsilon$

R is an achievable code if it is ε -achievable for all $0 < \varepsilon < 1$. T(X) denotes the minimum achievable source coding rate for X.

Def 8. Fix an integer $r \geq 2$. R is an achievable variable length source coding rate for X if for all $\gamma > 0$ and all sufficiently large n, there exists an r-ary prefix code such that the average codeword length L_n satisfies

$$(1/n)L_n \log r < R + \gamma$$

The minimum achievable variable length source coding rate for X is denoted by $\bar{T}(X)$

The following theorems of relevance to source coding have been proved in [5].

1. For any X and the Identity channel, $S(X) = T(X)$. The approach here is to consider a channel which essentially transmits the inputs unchanged (hence the identity channel) and thereby obtain the source coding result involving resolvability (a quantity originally defined for a channel W). Both this and subsequent theorems are proved by the method of proving both $S(X) \geq T(X)$ and $S(X) \leq T(X)$

2. For any X and the Identity channel, $\bar{S}(X) = \bar{T}(X)$
3. For any X and the identity channel, $S(X) = T(X) = \bar{I}(X; X)$. This result relates the resolution to the source coding rate to the sup entropy rate, thereby tying together two different information theoretic quantities and relating them to variable length source coding rate.
4. For every channel W and input process X , $S_\varepsilon(X) \leq \bar{I}(X; Y)$. This theorem relates the ε -resolvability (which is related to the source coding rate) to the sup information rate, analogous to the joint source channel coding theorem.
5. For any channel with finite input alphabet, $S_\varepsilon(X) \geq \sup_X \bar{I}(X; Y)$ for all $\varepsilon > 0$.
6. By the two previous theorems, we see that $S_\varepsilon(X) = \sup_X \bar{I}(X; Y)$ for all finite alphabet sources.
7. $\bar{S}_\varepsilon \leq \bar{S} \leq \bar{I}(X; Y)$ This result is identical to Theorem 4 except that resolvability has been replaced by mean resolvability.

For all of the above proofs, the criterion used was that of the vanishing variational distance. However, replacing this distance metric with the differential entropy does not make a significant difference as neither one is stronger than the other. In most general cases, bounds were obtained for the source coding rate in terms of resolvability. Only in the case of finite input alphabet, do we get equality of ε -resolvability and sup information rate. This is because, only in the finite alphabet case is it possible to obtain a lower bound on the distance between two outputs, given any two arbitrary inputs which are not identical (ref section IV lemma 6 in [5]). All of the above results viewed together constitute a concise and consistent set of results for source coding rates in terms of resolvability and related quantities which were previously not very widely used in information theory.

References:

- [1] B. McMillan, "The Basic Theorems of Information Theory," *Ann. Math. Stat* vol. 24, pp 196-219, 1953
- [2] P.C. Shields, "The Interactions Between Ergodic Theory and Information Theory," *IEEE Trans. Inform Theory*. Vol. 44, pp 2079- 2093, 1998
- [3] R.M. Gray and L.D. Davisson, "The Ergodic Decomposition of Discrete Random Processes," *IEEE Trans. Inform Theory*. Vol. IT-20, pp 625- 636, 1974
- [4] J.C. Kieffer, "Finite-State Adaptive Block to Variable Length Noiseless Coding of a Nonstationary Information Source," *IEEE Trans. Inform Theory*. Vol. 35, pp 1259- 1263, 1989
- [5] T.S. Han and S. Verdu, "Approximation Theory of Output Statistics," *IEEE Trans. Inform Theory*. Vol. 39, pp 752- 772, 1993
- [6] Y.G. Sinai, "Introduction to Ergodic Theory," Princeton University Press, 1976