Appl. Math. and Comp. Sci., 1995, vol.5, No.1, 139–168

OPERATIONAL RATE DISTORTION THEORY

ILAN SADEH*

The paper treats data compression from the viewpoint of information theory where a certain error probability is tolerable. We obtain bounds for the minimal rate given an error probability for blockcoding of general stationary ergodic sources. An application of the theory of large deviations provides numerical methods to compute for memoryless sources, the minimal compression rate given a tolerable error probability. Interesting connections between Cramer's functions and Shannon's theory for lossy coding are found.

1. Introduction

We study the problem of source coding with a fidelity criterion. The reader interested in an up-to-date and comprehensive survey of the subject can profitably consult the paper of Kieffer (1993). We consider the coding problem as a deterministic partition problem and obtain coding theorems for the general ergodic and stationary sources.

The source produces a random sequence $\{U_k\}$ and the decoder presents a random sequence $\{V_k\}$ to the user. In general, the finite alphabet V can differ from the source alphabet U. Given $u \in U$ and $v \in V$ a distortion-measure is any real positive function $d : [U \times V] \to \mathcal{R}^+$. The function d measures the distortion (cost, penalty, loss) suffered each time the source produces letters $u \in U$ and the user is presented with letters $v \in V$. Let $\rho_l(\bar{u}; \bar{v})$ – denote the distortion for a block- the average of the per letter distortions for the letters that comprise the block.

$$\rho_l(\bar{u}; \bar{v}) = \frac{1}{l} \sum_{i=1}^{l} d(\bar{u}_i; \bar{v}_i) \tag{1}$$

Let D be a given tolerable level of distortion relative to the memoryless distortion measure $d(u, v)$. Shannon's approach discusses the problem of D as a bound on the expected value of $\rho_l(\bar{u}; \bar{v})$. We will consider D as a bound on the average distortion. Both approaches converge as $l \to \infty$. The existence of blockcodes subject to a fidelity criterion has been proved under various assumptions on the class of sources, the fidelity criterion and the type of convergence of the rates to the rate distortion function $R(D)$. The problem was studied by Berger (1971), Ziv (1972), Neuhoff (1975), Kieffer (1978), Mackenthum and Pursley (1978), Ornstein and Shields (1990) and others.

* Department of Computer Sciences, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel, e-mail: sade@newton.bgu.ac.il

Our approach to the problem is different in the sense that we consider it mainly as a deterministic partition problem on a bipartite graph, unlike most of the known results which are based on random coding arguments.

We begin with the general ergodic stationary case. The basic idea is to study the subject by carving up the source output space into a high probability region, which can be partitioned so that all partition cells are contained in D-Balls of a radius D and the center of the D Ball is associated with a codeword (with zero distance) included in the Codebook. The low probability region, having probability P_e, contains the remaining vectors. The main results are obtained by using the partition and covering procedure. We present the tradeoffs among the the compression rate, the error probability and block length l for a given D. Our results, also apply, virtually without change in proof, to random fields.

Similar ideas have been used in practice by Eyuboglu and Forney (1993) to design lattice vector quantizers. A subset of a lattice covers the high probability region wherein every cell is inside a ball and a low probability region outside the lattice has no bound on distortion. By constraining P_e and then picking a good lattice, they have minimized the granular noise.

Next, a terse review of the Theory of Large Deviations, in particular the the asymptotic theory for Markov jump processes, described in (Knessl *et al.*, 1985), is given and followed by two examples in i.i.d. sources, the information and the empirical distribution. Next we apply the results of large deviations associated to these random vectors to treat the old problem of data compression for i.i.d. sources.

The problem for i.i.d. sources was first introduced by Shannon (1948). He returned to the problem in 1959 when he introduced the study of performance measures for data compression and provided the first source compression coding theorems (Shannon C.E. (1959). Strengthened versions of the coding theorems were proved by Berger (1971), Omura (1973), Ziv (1972), Blahut (1972); (1974) and others. Arimoto (1973), Dueck and Korner (1979) and Marton (1974) proved the exponential decay of the error probability for memoryless sources. We attempt to solve a question which has been open since Shannon's days. The issue is what is the expression of the minimal compression rate given a tolerable error probability. The solution is described by numerical methods, based on linear programming. The simplex structure plays a major part in the results. Interesting connections are found between Cramer's functions and Shannon theory.

The paper is organized as follows. Section 2 presents definitions and source compression coding theorems for the general ergodic sources. All are based on the partition and covering concept on bipartite graphs. Section 3 presents the general results of the asymptotic theory of large deviations, while section 4 presents the application to the empirical distribution and the average information of a process. Section 5 presents the results concerning data compression of i.i.d. (and Markovian) sources where all aspects of the subject are studied for finite blocklength. We treat in detail the issue of the best compression rate given a tolerable error probability. In section 6 we summarize the results and main contributions and outline future research trends in the area. The appendix presents in detail the asymptotic theory of large deviations.

2. The General Ergodic Stationary Case

The source produces a stationary ergodic sequence $\{U_k\}$ and the decoder presents a random sequence $\{V_k\}$ to the user. In general, the finite alphabet V can differ from the source alphabet U. The function d measures the distortion (cost, penalty, loss) suffered each time the source produces letters $u \in U$ and the user is presented with letters $v \in V$. Let $\rho_l(\bar{u}; \bar{v})$ – denote the distortion for a block- the average of the per letter distortions for the letters that comprise the block.

$$\rho_l(\bar{u}; \bar{v}) = \frac{1}{l} \sum_{i=1}^{l} d(\bar{u}_i; \bar{v}_i) \tag{2}$$

Let D be a given tolerable level of expected distortion relative to the memoryless distortion measure $d(u, v)$. The rate distortion function $R(D)$ is given by the minimal mutual information per source symbol subject to the constraint on the average distortion. It is known (Berger, 1971) that given a source u

$$\lim_{l \to \infty} \inf_{\widehat{Q}} \frac{1}{l} I_{\widehat{Q}}(u_0^{l-1}, v_0^{l-1}) = R(D) \tag{3}$$

where the infimum with respect to $\widehat{Q}$ is taken over all the conditional probability measures on $U^l \times V^l$ satisfying

$$E_{\widehat{Q}} \rho_l(\bar{u}, \bar{v}) \leq D \tag{4}$$

Shannon-Berger's theorem shows (Berger, 1971) that for stationary and ergodic sources, $R(D)$ is the lowest attainable rate by any block code with average distortion not exceeding D and it always exists. Gray *et al.* (1975) used process definitions to show that the minimization of the mutual information between input and output subject to an expected distortion constraint can be performed over stationary or ergodic processes.

The function $R(D)$ has been intensively studied, although much needs to be done in terms of obtaining explicit formulas for rate-distortion-function or, lacking this, obtaining iterative algorithms for computation of rate-distortion-functions. Some of its properties are:

1. It is convex U shape.
2. $R'(D)$ is continuous for $0 < D < D_{max}$.
3. $R'(D) \to -\infty$ as $D \to 0$.
4. $R'(D)$ is a monotonic non-decreasing function.

The value of $R(D)$ is the limit effective rate at which the source produces information subject to the requirement that the source output must be reproduced with an expected average distortion not exceeding the value D. However, it is always a limit value which is true only as l tends to infinity. In real life, we are interested in the rate convergence and the dependence on the blocklength l. Moreover, due to the nature of the problem, there are almost always sourcewords that cannot be encoded

to codewords in the Codebook. Hence, an error event occurs. We discuss the mutual dependence of compression ratio and the error probability.

The following definitions and theorems establish the relations in the general stationary ergodic case. We relate to the problem as a partition and covering on a bipartite graph where the sourcewords are located on one side of the graph and the codewords on the other side of the graph. A partition on the sourcewords space is performed subject to the fidelity criterion.

2.1. The Deterministic Partition Approach

The set of all possible codewords is partitioned into two disjoint subsets: Codebook and its complement set. The Codebook contains all the codewords in the code. Each sourceword $\bar{u}$ of length l is mapped onto exactly one of the codewords in the Codebook provided the distortion of the block is not larger than lD. Otherwise, the sourceword is included in the Error set and a coding failure is said to have occurred.

First we apply a deterministic partition to the sourcewords space and to the codewords. The partition algorithm assumes a fixed blocklength l and for a source with known probability structure $\boldsymbol{p} = p^1 \dots p^l$ defined on U^l. First we define a D-Ball covering on the sourcewords space.

Definition 1. A D-Ball covering of a codeword $\bar{v}$, denoted $\Upsilon(\bar{v})$, is a set of all sourcewords such that

$$\Upsilon(\bar{v}) = \left\{ \bar{u} \middle| \rho_l(\bar{u}, \bar{v}) \leq D \right\} \tag{5}$$

That is, we define spheres around all the possible codewords $\bar{v}$. But these spheres do not define probabilities on the codewords. Each sourceword should be mapped to exactly one codeword. Thus, we denote the set of the sourcewords that map to the codeword $\bar{v}$ after a partition, as $\boldsymbol{A}(\bar{v})$. We construct a partition such that for all $m \leq |V|^l$ the subsets $\boldsymbol{A}(\bar{v}^m)$ form a full partition of the set of all sourcewords. The probability of each codeword $\bar{v}$ is defined as the probability of the set $\boldsymbol{A}(\bar{v})$. Obviously, the l-order entropy of the codewords, denoted by $H_v(l)$ for all l, is also defined by the partition since the partition induces probabilities on the codewords, $\Pr(\bar{v}^j) = \Pr(\boldsymbol{A}(\bar{v}^j))$ for all j. The Codebook has the following properties as a sorted list of codewords,

$$\boldsymbol{A}(\bar{v}^j) \cap \boldsymbol{A}(\bar{v}^m) = \emptyset, \qquad \forall\, j \neq m$$

$$\Pr(\bar{v}^j) = \Pr\left(\boldsymbol{A}(\bar{v}^j)\right) \geq \Pr(\bar{v}^m) = \Pr\left(\boldsymbol{A}(\bar{v}^m)\right), \qquad \forall\, j < m$$

and consequently the induced l-order entropy is,

$$H_v(l) = -\frac{1}{l} E \log \Pr(\bar{v}).$$

Definition 2. An *acceptable partition* of blocklength l is a partition on the space of l length sourcewords such that for all $\bar{v}$, the associated subset $\boldsymbol{A}(\bar{v})$ satisfies $\boldsymbol{A}(\bar{v}) \subseteq \Upsilon(\bar{v})$ and that $\lim_{l\to\infty} H_v(l)$ exists.

As l tends to infinity we have to discuss a pair random process rather than finite words. If one has a pair random process $\{U_n, V_n\}$, then it will be of interest to find conditions under which there is a limiting per symbol distortion in the sense that

$$\rho_\infty(u, v) = \lim_{l \to \infty} \frac{1}{l} \rho_l(u^l, v^l)$$

exists. We require here that the distortion measure is single letter fidelity. It is shown by Gray (1990) that if the pair process is Asymptotically Mean Stationary (AMS) then the limiting distortion will exist and it is invariant from the ergodic theorem. We will restrict ourselves to stationary pair random process $\{U_n, V_n\}$.

By definition the induced entropy $H_v(l)$ for an acceptable partition algorithm tends to be the specific entropy rate H_v as obtained after a sequence of such partitions for blocklength l that tends to infinity. Actually, the partition is being performed on a space of realizations of stationary and ergodic process u that are mapped to stationary and ergodic processes v defined on a finite-valued alphabet V. However, we stress that we mean a partition which determines a deterministic mapping from the source process to the output process, as an extension of finite blocklength blockcoding, and it should be distinguished from an infinite sequence of coded blocks of finite length. A sequence of l length coded blocks is not stationary and clearly, its n order entropy, if the source has memory, does not converge to the entropy rate in the general case. The induced l-order entropy by using process definitions is obtained in the limit as,

$$\lim_{l \to \infty} H_v(l) = H_v \tag{6}$$

Definition 3. The set $D - \text{Ball}(\bar{u})$ is defined as,

$$D - \text{Ball}(\bar{u}) = \left\{ \bar{v} \middle| \rho_l(\bar{u}, \bar{v}) \leq D \right\} \tag{7}$$

Definition 4. The *operational rate distortion function* denoted by $\widehat{R}_l(D, P_e)$, is the minimal rate needed to cover with D-Balls a subset of sourcewords of probability $1 - P_e$.

Actually, Ornstein and Shields (1990) defined the limit operational rate distortion function as,

$$\widehat{R}(D) = \lim_{P_e \to 0} \lim_{l \to \infty} \widehat{R}_l(D, P_e) \tag{8}$$

That is, the best one can do on the average with block codes if an arbitrarily small part of the sourcewords space is removed.

Ornstein and Shields (1990) presented a blockcoding algorithm and proved almost sure convergence for Hamming distance. That is almost sure $\widehat{R}(D) = R(D)$. Their proof is true in principle for other distortion measure as well. Next, we show the extension of the celebrated Shannon McMillan Breiman Theorem (Breiman, 1957) to the lossy compression case. An error occurs in the event that a D-Ball around a sourceword $\bar{u}$ does not contain any word from the selected Codebook.

The following theorem generalizes that theorem (Breiman, 1957) by using the acceptable partition concept and the covering definition. Loosely, the theorems say that almost all the codewords sequences in the Codebook are nearly equiprobable and concentrated around the induced l-order entropy $H_v(l)$.

Block-Coding Theorem. *For any acceptable partition of blocklength l and given any $\delta > 0$, the set of all possible sourcewords of blocklength l produced by the source can be partitioned into two sets, Error and Error^c, for which the following statements hold:*

i) *Assuming a stationary system, the probability of a sourceword belonging to Error, vanishes as l tends to infinity.*

ii) *If a sourceword $\bar{u}$ is in Error^c, then its associated codeword $\bar{v}$ is in the Codebook and its probability of occurrence is more than $e^{-l(H_v(l)+\delta)}$.*

iii) *The number of codewords in the Codebook is at most $e^{l(H_v(l)+\delta)}$.*

Proof. The idea of the proof is based on optimal selection of codewords for the Codebook. Optimality means minimum error probability for a given acceptable partition. Once the acceptable partition on the sourcewords space is determined and performed, the best selection of the Codebook in the sense of minimal error probability, is to take for the Codebook all the most probable codewords as produced by the acceptable partition algorithm.

The partition algorithm induces the probabilities that are assigned to each codeword. Recall that

$$-\frac{1}{l}E\log\Pr(\bar{v}) = H_v(l) \tag{9}$$

We define a certain threshold probability p_t, and define the Codebook Set by,

$$\Gamma_l = \left\{\bar{v}|\,\Pr(\bar{v}) \geq p_t\right\} \tag{10}$$

Its cardinality is denoted $\mid \Gamma_l \mid$. We choose p_t such that $p_t = e^{-l(H_v(l)+\delta)}$ where $H_v(l)$ is the l order entropy of the process v as determined by the l length partition algorithm. Thus, the requirements of Statement 2 of the theorem are satisfied and the Codebook set is defined by,

$$\Gamma_l(D,\delta) = \left\{\bar{v}|\,\Pr(\bar{v}) \geq \exp\left(-l(H_v(l)+\delta)\right)\right\} \tag{11}$$

Clearly, $\mid \Gamma_l(\delta, D) \mid e^{-l(H_v(l)+\delta)} \leq 1$. Thus Statement 3 of the theorem holds.

An error occurs in the event that a D–Ball around a sourceword $\bar{u}$ does not contain any word from the selected codebook. Thus, the Error set is obtained by using (10),

$$Error(\delta, D) = \left\{\bar{u}|\quad \max_{\bar{v}:\rho(\bar{u},\bar{v})\leq D} \Pr(\bar{v}) < p_t\right\} \tag{12}$$

after substituting for p_t and using (11) we have,

$$Error(\delta, D) = \left\{ \bar{u} | \quad \min_{\bar{v}:\rho(\bar{u},\bar{v})\leq D} -\frac{1}{l}\log \Pr(\bar{v}) - H_v(l) \geq \delta \right\} \tag{13}$$

The probability of the codewords is induced by the partition on the sourcewords space. Thus, the error probability is equal to the probability of the compliment set of the C codebook. That is,

$$\Pr\left\{ Error(\delta, D) \right\} = \Pr\left\{ \bar{v} | \quad -\frac{1}{l}\log \Pr(\bar{v}) - H_v(l) \geq \delta \right\} \tag{14}$$

The system preserves stationarity. Hence, as we increase the blocklength l to infinity, we map each source process to an ergodic and stationary process. For a given source process u, the output v is a finite-valued stationary ergodic process with entropy rate H_v. We use the Breiman refinement of the Shannon–McMillan theorem (Breiman, 1957) and its modern modifications of Barron (1985), Orey (1985), Algoet and Cover (1988), which assert that,

$$\Pr\left\{ \lim_{l\to\infty} -\frac{1}{l}\log \Pr(\bar{v}) = H_v \right\} = 1 \tag{15}$$

The convergence almost surely to the entropy rate implies that the Error set (12) has a measure that tends to zero for all positive δ as $l \to \infty$. This completes the proof of Statement 1. ■

Actually, the theorem can be generalized to AMS systems and additive fidelity criteria. A large majority of information theory is devoted to additive distortion measures and this bias is reflected in this work. Such generalization uses the theory of Gray (1990) and the strong Asymptotic Equipartition Property (AEP) that was recently proved by Algoet and Cover (1988) for processes that are stationary but not necessarily ergodic and for AMS processes satisfying an extra hypothesis.

We can obtain information from the theorem on the tradeoffs among Codebook size and error probability, given D. The error event contains information measured by $-\log P_e$. Consider the information related term of the problem, δ, as the original parameter of the problem. That is, given codewords of length l and δ extra nats of information, we will try to estimate the number of nats (bits) needed for D-semi-faithful encoding of a symbol from the sourcewords whose associated codeword's average information does not exceed $H_v(l)+\delta$ *nats*. It is obvious that $H_v(l)+\delta$ *nats* would be enough, since any of the considered codewords has lower average information. We conclude from the theorem that the number of *nats* needed is at most $H_v(l) + \delta$. We expect to gain a few *nats* from the non-uniform distribution of codewords in the Codebook. Most of the mass is obviously concentrated around $H_v(l)$ *nats*. Next, we state the properties of the Codebook where the error probability P_e, the distortion level D and the blocklength l are given.

Corollary. *Given is a stationary ergodic source u with known probabilities for all blocklengths l, an acceptable average distortion D and a tolerable error probability*

P_e. Assuming the l order entropy induced by the chosen acceptable partition is $H_v(l)$, then the following statements hold,

i) *The optimal code set is,*

$$\Gamma_l(D,\delta) = \left\{ \bar{v} \middle| \Pr(\bar{v}) \geq e^{-l(H_v(l)+\delta)} \right\} \tag{16}$$

where a value δ is determined by the error probability.

ii) *The error set is defined by,*

$$Error_l(\delta, D) = \left\{ \bar{u} \middle| \min_{\bar{v}:\rho(\bar{u},\bar{v})\leq D} -\frac{1}{l}\log \Pr(\bar{v}) - H_v(l) > \delta \right\} \tag{17}$$

3. An Asymptotic Theory of Large Deviations

A summary of the results of (Knessl *et al.*, 1985) and of Schuss–Gofman unpublished work is presented. Let $\{X_n\}$ be K-dimensional mean of i.i.d. Z_n such that,

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

where $\{Z_n\}$ is a zero mean K-dimensional i.i.d. whose p.d.f. is

$$\frac{\partial^K}{\partial z_1 \cdots \partial z_K} \Pr\left\{ Z_n \leq z \right\} = w(z) \tag{18}$$

We are interested in the evaluation of the probability density function, for large n:

$$p(\bar{y}, n) = \frac{\partial^K}{\partial y_1 \cdots \partial y_K} \Pr\left\{ X_n \leq \bar{y} \right\} \tag{19}$$

We generalize the main results of the one dimensional theory of (Knessl *et al.*, 1985) to higher dimensions. The principles are described in the Appendix. Following (Knessl *et al.*, 1985) Method we construct an approximation to $p(\bar{y}, n)$ for $n \gg 1$.

$$p(\bar{y}, n) \sim \sqrt{n}^K \exp\Big(-n\Psi(\bar{y})\Big)\Big(K_0(\bar{y}) + K_1(\bar{y})/n + \cdots\Big) \tag{20}$$

$M(\theta)$ is the moment generating function of Z_n

$$M(\theta) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[\exp(\theta z) w(z)\Big]\, \mathrm{d}z_1 \cdots \mathrm{d}z_K \tag{21a}$$

It is known that

$$\Psi(\bar{y}) = \bar{y}\,\nabla\Psi(\bar{y}) - \log M\Big(\nabla\Psi(\bar{y})\Big) \tag{21b}$$

which is equivalent to:

$$\Psi(\bar{y}) = \sup_{\theta}\Big(\theta\bar{y} - \log M(\theta)\Big) \tag{21c}$$

It is sufficient to compute $K_0(\bar{y})$ for almost all practical applications. We show that the approximated density function is:

$$p(\bar{y}, n) \approx (2\pi/n)^{-K/2}\sqrt{\det C(\bar{y})}\exp\left\{-n\Psi(\bar{y})\right\} \tag{22}$$

where

$$C_{i,j}(\bar{y}) = \frac{\partial^2\Psi(\bar{y})}{\partial y_i \partial y_j} \tag{23}$$

4. The Empirical Distribution and The Information

We apply the large deviations results of Section 3 to the memoryless source case.

4.1. The Empirical Distribution

The probability $p(u)$ is the *a priori* probability of a source letter u taken from the source alphabet U. Denote by $\bar{u}$ a word in U^l.

Definition 6. $\mathcal{N}(u;\bar{u})$ denotes the number of indices $k \in [1 \cdots l]$ such that $\bar{u}_k = u$. That is, the number of occurrences of the letter u in the vector $\bar{u} = \{\bar{u}_1 \cdots \bar{u}_l\}$.

Definition 7. The random vector $X_l(\bar{u})$ defines the deviations of the empirical distribution from the probability measure p.

$$X_l(\bar{u}) = \left\{\left(\frac{\mathcal{N}(u_1,\bar{u})}{l} - p(u_1)\right); \ldots; \left(\frac{\mathcal{N}(u_{N-1},\bar{u})}{l} - p(u_{N-1})\right)\right\} \tag{24}$$

with $K = N - 1 = |U| - 1$ components.

For a stationary and memoryless source – the random vector $X_l(\bar{u})$ is a large deviations process with $K = N - 1$ components, satisfying $X_l = \frac{1}{l}\sum_{i=1}^{l} Z_i$ where the random vector Z_l is a sequence of zero mean, i.i.d. K-dim random variable, such that $\Pr\left(Z_l = \bar{e}_i - \bar{p}\right) = p_i$. Thus, following (18), its probability density function $w(z)$ is defined by

$$w(z) = \Pr\left\{Z_l = z\right\} = \sum_{i=1}^{K} p_i \delta(\bar{e}_i - \bar{p} - \bar{z}) \tag{25}$$

where $\bar{e}_i$ is a unit vector at the i-th index and $\bar{p}$ is the vector of the *a priori* probabilities.

The result can be applied in a straightforward manner to the jointly pair or triples sequences and so on. We define the deviation from the joint distribution,

$$X_l(\bar{u},\bar{v}) = \left\{\left(\frac{\mathcal{N}(u_1,v_1;\bar{u},\bar{v})}{l} - p(u_1,v_1)\right) \ldots \left(\frac{\mathcal{N}(u_N,v_{M-1};\bar{u},\bar{v})}{l} - p(u_N,v_{M-1})\right)\right\}$$

$$\forall\,(u,v) \in U \times V - \{u_N, v_M\}$$

where $\bar{p} = \Big(p(u_1, v_1) \ldots p(u_N, v_{M-1})\Big)$. Following (18) we obtain the density function,

$$w(\bar{z}) = \sum_{k=1}^{N*M-1=K} p_k \delta(\bar{e}_k - \bar{p} - \bar{z})$$

Without loss of generality we have omitted the last component (N, M) to construct $K = N*M-1$ random variables. Obviously, the K components are mutually dependent. This vector is defined in the sample space $\mathcal{R}^K$.

4.2. The Information

Definition 8. Given a memoryless source with entropy H and sequences of block-length l, we define the deviation of the information from entropy as the i.i.d. random variables $\tau_i(u_i) = -\log p(u_i) - H \quad i = 1 \ldots l$.

The sample mean of the deviations from the entropy is denoted by

$$X_l^w(\bar{u}) = \frac{1}{l} \sum_{i=1}^{l} \Big(-\log p(u_i) - H \Big) \tag{26a}$$

It is clear that $E(X_l^w) = 0$.

Following (18), the probability density function of the deviation of the average information is:

$$w(z) = \sum_{i=1}^{|U|} p_i \delta(-\log p_i - H - z) \tag{26b}$$

4.3. The Mutual Connections

We show that the random variable associated with the weakly typical sequences (the information), can be easily derived from the distribution of the empirical distribution for memoryless sources. That is, because

$$\Pr(\bar{u}) = \prod_1^N p(u_1)^{\mathcal{N}(u_1, \bar{u})} \cdots p(u_N)^{\mathcal{N}(u_N, \bar{u})}$$

we have from (26a) that,

$$X_l^w(\bar{u}) = -\sum_1^N \log p(u_1) \Big(\frac{\mathcal{N}(u_1, \bar{u})}{l} - p(u_1) \Big) + \ldots$$

$$+ \log p(u_N) \Big(\frac{\mathcal{N}(u_N, \bar{u})}{l} - p(u_N) \Big)$$

Denoting the vector of constants,

$$\bar{f}_i = \log \frac{p(u_i)}{p(u_N)}, \qquad i = 1 \ldots K = N - 1 \tag{27}$$

we find that

$$X_l^w(\bar{u}) = -X_l(\bar{u})\bar{f} \tag{28}$$

4.4. The Distribution of the Empirical Distribution

We present here the appropriate terms in the general large deviations distribution function, that identify the distribution of the deviations of the empirical distribution. Following (22),

$$\begin{aligned} p(\bar{y}, l) &= \frac{\partial^K}{\partial y_1 \cdots \partial y_K} \Pr\left\{X_l(\bar{u}) \leq \bar{y}\right\} \\ &= (2\pi/l)^{-K/2} \sqrt{\det C(\bar{y})} \exp\left\{-l\Psi(\bar{y})\right\} \end{aligned} \tag{29a}$$

where:

$$\Psi(\bar{y}) = \sup_{\bar{\theta}} \left(\bar{\theta}\bar{y} - \log M(\bar{\theta})\right)$$

Setting $\bar{p} = \left(p(u_1) \ldots p(u_{N-1})\right)$ we have,

$$M(\bar{\theta}) = \exp(-\bar{\theta}\bar{p}) \sum_{k=1}^{N-1=K} p_k \exp(\bar{\theta}_k)$$

$$\frac{\partial \Psi(\bar{y})}{\partial y_k} = \bar{\theta}_k = \log\left(\frac{p_k + \bar{y}_k}{p_k}\right)$$

where $\Psi(0) = 0$. Hence

$$\Psi(\bar{y}) = \sum_{k=1}^{K} \log\left(\frac{p_k + \bar{y}_k}{p_k}\right)(\bar{y}_k + p_k) - \log\left(\frac{\sum_{k=1}^{K}(p_k + \bar{y}_k)}{\sum_{k=1}^{K} p_k}\right) \tag{29b}$$

$$C_{i,j}(\bar{y}) = \frac{\partial^2 \Psi(\bar{y})}{\partial y_i \partial y_j} = \left(\frac{1}{p_i + \bar{y}_i}\right)\delta_{i,j}$$

$$\det C(\bar{y}) = \prod_{k=1}^{K} \left(\frac{1}{p_k + \bar{y}_k}\right). \tag{29c}$$

4.5. The Distribution of the Information

The p.d.f. of the deviation of the sample mean of the information from the entropy is given following (22) by,

$$\begin{aligned} p(y, l) &= \frac{\mathrm{d}}{\mathrm{d}y} \Pr\left\{X_l^w(\bar{u}) \leq y\right\} \\ &\approx \left(\frac{l}{2\pi}\right)^{1/2} \sqrt{\psi''(y)} \exp\left\{-l\psi(y)\right\} \end{aligned} \tag{30a}$$

where:

$$\psi(y) = \sup_{\theta}\Big(\theta y - \log M(\theta)\Big)$$

which is equivalent to the solution of

$$\psi(y) = y\psi'(y) - \log M\Big(\psi'(y)\Big) \tag{30b}$$

$$\psi(y) = \sup_{\theta}\left(\theta(H+y) - \log\sum_{i=1}^{|U|} p(u_i)^{1-\theta}\right) \tag{30c}$$

5. Coding of Memoryless Sources

The requirement for data compression comes when one needs to communicate source data that is generated at a rate greater than channel capacity. We address the following issue: given a tolerable probability of error P_e what is the best compression ratio R for a finite blocklength l and distortion level D. Such issues arise in the design of communications systems, such as image communications, where channel capacity and source statistics are known and where a specific rate of error is tolerable. The exact connections between all these parameters are studied for discrete memoryless sources, and can be extended to more complicated cases. Random fields can be analyzed in a similar way.

Recent research trends have been in parallel directions. The first is a determination of $R(D)$. If the source is memoryless, then explicit formulas are known for certain distortion measures (Berger, 1971 – ch.2), and for general distortion measures the algorithm of Blahut (1972) can be used to compute successive approximations to $R(D)$. However, for more complicated cases, a description of $R(D)$ valid for all D is not known. On the other hand, error exponents in source coding have been studied intensively. Considerable success has been achieved in determining the nature of error exponents in the zero-error case (i.e., the $D = 0$ case) when the source is memoryless or Markov source. Results were obtained in Anantharam (1990), Csiszar and Longo (1971), Davisson *et al.* (1981), Longo and Sgarro (1979), Merhav and Neuhoff (1992), Natarajan (1985) and Covo and Schuss (1991). Much less is known concerning error exponents in the $D > 0$ case. Marton (1974) obtained a result for a memoryless source. In this section we will unify the two approaches and discuss the combination of the two issues. That is, for a given P_e, a bound on the average distortion level D and a blocklength l, we shall seek the best compression rate. The results, developed for memoryless source might be generalized for classes of sources for which there is a well-developed body of large deviations results for the source output process. (See also applications of large deviations theory to source coding by Sadowsky and Bucklew (1990).)

Our approach to the problem is described in the following steps;

a. Definitions and explanation of the stochasic approach.

b. Transformation of the deterministic problem to a stochastic one and calculation of the error probability.

c. Performance analysis and procedures to calculate the rate R for i.i.d. (and Markovian) sources, by specifying the tolerable error probability.

5.1. Definitions and the Stochastic Approach

Given is a source u with known probabilities and an acceptable average distortion D. Then for each blocklength l, a partition algorithm induces the probabilities on the codewords and define entropy $H_v(l)$. We omit the subscript l for convenience. The encoding procedure defines the optimal code set,

$$\Gamma_l(D, \delta, H_v) = \left\{ \bar{v} | \Pr(\bar{v}) \geq \exp\left(-l(H_v + \delta)\right) \right\} \tag{31}$$

The value δ is determined by the tolerable error probability. We describe the Error set as a function of the induced entropy H_v and *del*.

$$Error(\delta, D, H_v) = \left\{ \bar{u} | \min_{\bar{v}:\rho(\bar{u},\bar{v})\leq D} -\frac{1}{l} \log \Pr(\bar{v}) - H_v \geq \delta \right\} \tag{32}$$

As mentioned above the encoder- decoder pair is a deterministic machine. Nevertheless, when restricting our attention to a single source output symbol without knowledge of the previous output symbols, the reproducing symbol is not predetermined. At this level it is an output of a transition matrix and a memoryless source, even though at the block level the encoding is deterministic. We may describe the deterministic coding algorithm as a stochastic one with the "best" transition matrix Q that simulates the data compression at the level of a single symbol. By considering the deterministic problem as a stochastic one, we use the theory of large deviations and describe all induced properties as a function of Q.

Hereandafter the index i denotes a letter in the source alphabet and j denotes the reproducing letters.

Definition 9. The entropy of codewords H_v of a memoryless source is given by,

$$H_v(Q) = -\sum_j \sum_i p(i)q(j|i) \log \sum_i p(i)q(j|i) \tag{33}$$

Definition 10. The expected value of the distortion between the sourcewords and codewords, denoted as d_0, is a function of the transition matrix Q, and its value is,

$$d_0(Q) = \sum_j \sum_i p(i)q(j|i)d(i,j) \tag{34}$$

Definition 11. The vector $\bar{x}$ with $N-1$ components denotes the deviations of the empirical distribution of a sourceword, from the probability measure defined on the alphabet U. That is,

$$\bar{x}_{k=i} = \left(\frac{\mathcal{N}(i|\bar{u})}{l} - p(i) \right) \tag{35}$$

Definition 12. The vector $\bar{y}$, with $NM-1$ components, denotes the deviations of the joint empirical distribution of one sourceword and a codeword, from the joint probability measure defined on the Cartesian product $U \times V$,

$$\bar{y}_{k=(i,j)} = \left(\frac{\mathcal{N}(i,j|\bar{u},\bar{v})}{l} - p(i,j) \right) \tag{36}$$

Definition 13. The vector $\bar{d}$ with $NM-1$ components, is defined by

$$\bar{d}_{k=(i,j)} = \Big(d(i,j) - d(N,M) \Big) \tag{37}$$

Without loss of generality we may assume that $d(N,M)=0$.

Definition 14. The vector $\bar{t}$ is defined as]

$$\bar{t}_{k=(i,j)} = \log\left(\frac{p(j)}{p(M)} \right) \tag{38}$$

Without loss of generality we may assume that $p(M)$ is the least probable symbol with positive probability.

Recall the definitions of $\Psi(\bar{x}), C(\bar{x})$ in (29a)–(29c) and $\psi(y)$ in (30a)–(30c).

5.2. The Operational Rate Distortion Function

In this section we deal with the analysis of the classical problem of D-semifaithful source coding known as data compression of a memoryless source with dependence on the algorithm simulated by Q. We provide numerical procedures to estimate the best compression given a tolerable error probability P_e.

The Operational Rate Distortion Theorem. *The minimal compression rate R given a tolerable error probability P_e, a distortion level D, by blockcoding of blocklength l for an i.i.d. source, is given by*

$$R \approx \inf_{Q} \left\{ H_v + \delta - \psi(\delta) - \frac{\log l}{2l} + \frac{\log \sqrt{\frac{\psi''(\delta)}{(2\pi)}} \frac{1}{(1-\psi'(\delta))}}{l} \right\} \tag{39}$$

where Q is found iteratively and δ is determined iteratively by the tolerable error probability P_e.

The minimum of $\Psi(\bar{x})$ on the boundary of $\mathcal{D}$ defined in (41) is given by

$$-\frac{\log P_e}{l} \approx \Psi(\bar{x}^*) = \inf_{\bar{x} \in \partial\mathcal{D}} \Psi(\bar{x}) \tag{40}$$

The term δ defines the region $\mathcal{D}$ for the matrix Q such that (40) is satisfied and,

$$\mathcal{D} = \left\{ \bar{x} \mid \bar{y}^*(\bar{x}, D)\bar{t} > \delta \right\} \tag{41}$$

where $\bar{y}^*(\bar{x}, D)\bar{t}$ is the solution of the following simplex problem,

$$\begin{aligned} &\textit{Minimize} \quad \bar{y}\bar{t} \quad \textit{subject to} \\ &\mathcal{A}\bar{y} = \bar{x} \\ &-p(u_i, v_j) \leq \bar{y}\bar{e}_{k=(i,j)} \leq 1 - p(u_i, v_j) \quad \forall\, k = 1 \ldots NM - 1 \\ &p(u_N, v_M) - 1 \leq \bar{y}\bar{1}_y \leq p(u_N, v_M) \\ &\bar{y}\bar{d} \leq D - d_0 \end{aligned} \tag{42}$$

where $\mathcal{A}$ represents the transform matrix between $\bar{y}$ to $\bar{x}$, $\bar{1}_y$ is a vector with 1 in all of its components and $\bar{e}_k$ is a unit vector. The terms ψ, Ψ (29a)–(29c), (30a)–(30c), and $d_0, p(u, v), \bar{t}, H_v$ are known functions of Q (33)–(38).

The Markovian source case has a similar expansion.

Proof.

STEP 1: The Error Event.

Define the random variable as in (17),

$$X_l^w(\bar{v}) = -\frac{1}{l}\log \Pr(\bar{v}) - H_v \tag{43}$$

Following (31) we choose the code set to be all the codewords that satisfy the condition $\Pr(X_l^w(\bar{v}) \leq \delta)$ for some transition matrix Q which simulate the partition that induce probabilities on the codewords.

Given the partition simulated by Q, we define the random variable $P(\bar{u})$ as the probability of the most probable codeword contained in the D-Ball around the sourceword $\bar{u}$.

$$P(\bar{u}) = \max_{\bar{v}:\rho(\bar{u},\bar{v})\leq D} \Pr(\bar{v}) \tag{44}$$

and from (1) and Definition 6, the average of the per letter distortions for the letters that comprise the l length blocks is,

$$\rho_l(\bar{u}, \bar{v}) = \sum_j \sum_i \frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l} d(i, j) \tag{45}$$

Hence, (44) is transformed to,

$$P(\bar{u}) = \max_{\bar{v}:\sum_j \sum_i \frac{\mathcal{N}(u_i,v_j|\bar{u},\bar{v})}{l} d(i,j)\leq D} \prod_{j=1}^{M} p(v_j)^{\sum_i \mathcal{N}(u_i,v_j|\bar{u},\bar{v})} \tag{46}$$

The error is the event that a D-Ball around the sourceword $\bar{u}$ does not contain any word from the selected codebook defined in (31). Following (32) (44), recall the definition of the error event,

$$Error(\delta, D) = \left\{ \bar{u}| \; -\frac{1}{l}\log P(\bar{u}) - H_v > \delta \right\} \tag{47}$$

Next, we define the random variable $Z(\bar{u})$ from (44) and (47),

$$Z(\bar{u}) = -\frac{1}{l}\log P(\bar{u}) - H_v \tag{48}$$

which by (46) is,

$$Z(\bar{u}) = \min \sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l} - p(u_i, v_j)\Big)\log p(v_j)$$

subject to

$$\forall\, i = 1 \ldots N-1 \qquad \sum_{j=1}^{M}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big) = \Big(\frac{\mathcal{N}(u_i|\bar{u})}{l}\Big)$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big) = 1$$

$$\forall\, i, j \quad 1 \geq \frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l} \geq 0$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big)d(i, j) \leq D$$

or

$$Z(\bar{u}) = \min \sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l} - p(u_i, v_j)\Big)\log p(v_j)$$

subject to

$$\begin{aligned}
&\forall\, i = 1 \ldots N-1 \qquad \sum_{j=1}^{M}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big) - p(u_i, v_j) = \Big(\frac{\mathcal{N}(u_i|\bar{u})}{l}\Big) - p(u_i)\\
&\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big) - p(u_i, v_j) = 0\\
&\forall\, i, j \quad 1 \geq \frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l} \geq 0\\
&\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{\mathcal{N}(u_i, v_j|\bar{u}, \bar{v})}{l}\Big) - p(u_i, v_j)d(i, j) \leq D - d_0
\end{aligned} \tag{49}$$

Without loss of generality we may assume that $p(M)$ is the least probable code symbol. We denote by the index N the source symbol with zero distance from the

least probable code symbol M i.e. $d(N, M) = 0$. Therefore, all elements of $\bar{t}$ and all elements of $\bar{d}$ are non negative.

Thus using Definitions 9–14 formula (49) changes to the form of a simplex problem,

$$Z(\bar{x}) = \min\ \bar{y}\bar{t}$$

subject to

$$\begin{aligned} &\mathcal{A}\bar{y} = \bar{x} \qquad (50)\\ &-p(u_i, v_j) \leq \bar{y}\bar{e}_{k=(i,j)} \leq 1 - p(u_i, v_j) \quad \forall\, k = 1 \ldots NM - 1\\ &p(u_N, v_M) - 1 \leq \bar{y}\bar{1}_y \leq p(u_N, v_M)\\ &\bar{y}\bar{d} \leq D - d_0 \end{aligned}$$

where $\mathcal{A}$ represents the transform matrix between $\bar{y}$ to $\bar{x}$, $\bar{1}_y$ is a vector with 1 in all of its components , and $\bar{e}_k$ is a unit vector.

We denote the vertex point $\bar{y}^*(\bar{x}, D)$ as the minimum feasible solution which minimizes $\bar{y}\bar{t}$ under the constraints. Thus,

$$Z(\bar{x}) = \bar{y}^*(\bar{x}, D)\bar{t} \qquad (51)$$

Next, we identify the region of $\bar{x}$ that correspond to the error event in (46)–(47) and the definition of $\bar{x}$ in Definition 11,

$$\mathcal{D} = \left\{ \bar{x} \mid \bar{y}^*(\bar{x}, D)\bar{t} > \delta \right\} \qquad (52)$$

Note that the deviations of the empirical distribution of the sourceword $\bar{u}$ denoted by $\bar{x}$ are defined as a random vector variable depending on $\bar{u}$. Thus, we can describe the error event with $\bar{x}$ instead of $\bar{u}$ as in (32). The probability of the error event, following (47)–(52), is

$$\Pr\left(Z(\bar{x}) > \delta\right) = \int_{\mathcal{D}} p(\bar{x}, l)\, \mathrm{d}\bar{x} \qquad (53)$$

where $p(\bar{x}, l)$ is the p.d.f. of the deviations of the empirical distribution of sourcewords from the probability measure, as given in (29a)–(29c). The minimum error probability is obtained by

$$P_e(\delta, l, D) = \int_{\mathcal{D}} p(\bar{x}, l)\, \mathrm{d}\bar{x} \qquad (54)$$

where the term δ determines the size of the codebook as $\boldsymbol{S} = \exp lR$.

The value of $P_e(\delta, D, l)$ (54) can be evaluated by the large deviations theory (18)–(30). The error probability is found in regions where integral (53) converges to its approximate value.

In the region where the origin $\bar{x} = 0$ is included in $\mathcal{D}$, see (52), integral (53) is almost equal to 1, while in all other regions of (D, δ) integral (53) decays exponentially as l increases.

Suppose the zero point $\bar{x} = 0$ is not an interior point of $\mathcal{D}$. Then, the minimum of $\Psi(\bar{x})$ on the boundary of $\mathcal{D}$ is obtained at $\bar{x}^*$. That is,

$$\Psi(\bar{x}^*) = \inf_{\bar{x} \in \partial\mathcal{D}} \Psi(\bar{x}) \tag{55}$$

Clearly, the point $\bar{x}^*$ is a function of (δ, D, l). Hence, the minimal probability of error is derived from (52)–(55), and has the form

$$P_e(\delta, D, l) \approx \begin{cases} 1 & \text{if } (\bar{x} = 0) \in \mathcal{D} \\ \sqrt{\left(\frac{l}{2\pi}\right)^K \det C(\bar{x}^*)} \exp\left\{ - l\Psi(\bar{x}^*)\right\} & \text{if } (\bar{x} = 0) \in \mathcal{D}^c \end{cases} \tag{56}$$

where $\mathcal{D}$ is in (52).

STEP 2: Finding the rate as a function of δ.

In this step we consider the construction of the Codebook, as defined in (31). The error event is equivalent to the occurrence of a codeword, after the optimal partition has induced the probabilities on the codewords, outside the Codebook Γ_l. That event is also equivalent to a large deviation of $X_l^w(\bar{v})$ from its mean- the entropy $H_v(l)$. In other words, once the partition simulated by Q has induced probabilities and entropy on the codewords, an error occurs upon the appearance of a codeword containing more than an average of $H_v + \delta$ nats per symbol. The δ nats of excess information allowed per symbol are related to the coding rate R. In this step we still have to find the functional relation between δ, l and R. We use the density function of the deviation of average information from entropy as found in (30a)–(30c) for that goal.

As in (Covo and Schuss, 1991), we slice the set $\Gamma_l(\delta)$ (see (31)), into disjoint subsets each containing words $\bar{v}$ with almost equal probabilities. Let T_m be such a subset, using (43);

$$T_m(y_m, y_{m+1}) = \left\{ \bar{v} \mid y_m < X_n^w(\bar{v}) \leq y_{m+1} \right\}$$

where $-H_v = y_0 \leq y_1 \ldots \leq y_{max} = \delta$. Since all $\bar{v} \in T_m$ have equal probabilities, we have that the number of words in each slice is nearly the slice probability divided by the probability of each member. That is,

$$|T_m| = p(y_m, l)\Delta y_m \exp l(H_v + y_m)$$

The set $\Gamma_l(\delta)$ is the union of all the distinct subsets T_m. Hence, summing over the slices gives the size of Codebook.

$$\mid \Gamma(\delta) \mid = \sum_m |T_m| = \sum_m p(y_m, l)\Delta y_m \exp l(H_v + y_m) \tag{57}$$

By the mean value theorem and as the partition-norm is decreasing to zero — the summation (57) tends to

$$\mid \Gamma(\delta) \mid = \int_{-H_v}^{\delta} \exp n(H_v + y) p(y, l)\, \mathrm{d}y \tag{58}$$

where $p(y, l)$ is the p.d.f. of the deviation of the average information of the codewords $\bar{v}$ from their entropy given in (30a)–(30c).

On the other hand, we have defined the cardinality of the code as

$$\mid \Gamma_l(\delta) \mid = \exp(lR) \tag{59}$$

Thus, by equating both expressions for $\mid \Gamma(\delta) \mid$, (58) and (59), we obtain

$$e^{lR} = \int_{-H_v}^{\delta} \sqrt{\frac{l\psi''(y)}{2\pi}} e^{(l(H_v + y - \psi(y)))}\, \mathrm{d}y \tag{60}$$

In (Covo, 1992) it is shown that the exponent function $y - \psi(y)$ reaches its maximum in the relevant range at the boundary point $y = \delta$. Therefore we expand $y - \psi(y)$ in Taylor's series about δ. Then we change variables and use Laplace method to evaluate integral (60). It is known that the value of the integral for large l is affected only by the neighborhood of δ (Bleistein and Handelsman, 1975). That is,

$$e^{lR} = \sqrt{\frac{\psi''(\delta)}{2\pi l}} \frac{1}{1 - \psi'(\delta)} e^{(l(H_v + \delta - \psi(\delta)))} \left(1 + \mathcal{O}\left(\frac{1}{l}\right)\right) \tag{61}$$

after neglecting terms of $O(1/l)$ in the exponent, and terms which decay exponentially. The relation between R, l, δ is found from (30a)–(30c), and (61), as

$$R \approx \left(H_v + \delta - \psi(\delta)\right) - \frac{\log l}{2l} + \frac{\log \sqrt{\frac{\psi''(\delta)}{(2\pi)}} \frac{1}{(1 - \psi'(\delta))}}{l} \tag{62}$$

We have omitted most of the technical development and its reasoning, since we believe it does not contribute to the understanding of the whole matter. All details are presented in (Covo, 1992), where the lossless encoding is treated in details.

STEP 3: Optimization.

To conclude the proof, we take the infimum over the set of all matrices Q that satisfy the expression for P_e (50)–(56), and substitute that Q in the expression for R (62). ■

Now we may conclude the above result by the following "law". The transition matrix Q that simulates the scheme which minimizes the compression rate for a given error probability, is chosen from the set of transition matrices that satisfy (40)–(42) such that the expression $H_v(Q) + \delta - \psi_Q(\delta, l)$, is minimized. The loss of $\Psi_Q(\delta)$ amount of information in the transmission results in the compression by gaining $\psi_Q(\delta)$ nats. The term δ is determined by the tolerable error probability. We obtain a "conservation law", where the amount of the lost information is equal to the gain in the

compression, only in the lossless case. It is an interesting interpretation for the two Cramer's functions in context of lossy data compression.

6. Summary

Major contributions of this paper include:

i) Using the concept of *an acceptable partition* on a bipartite graph where the partition is carried on the sourcewords space on one side of the graph towards the codewords space on the other side. The encoding procedure is defined by covering the partitioned sourcewords space by the codewords. The approach is different than the usual source- compression coding theorems based on Shannon's theory and random coding arguments. A possible extension to the problem of finding a good lattice vector quantizer is currently receiving much attention.

ii) The properties of the optimal codebook given a tolerable error probability P_e, a distortion level D and a blocklength l.

iii) The issue of transmission under tolerable uncertainty is addressed. That is, the iterative algorithm for computation of the minimal compression rate given a tolerable error probability. We are able by using exact numerical methods based on linear programming, to give the minimal attainable rate given a tolerable error probability.

iv) Interesting connections between Cramer's functions ψ (of output information), Ψ (of joint empirical distribution), and the Shannon theory are found. The transition matrix Q that simulates the compression scheme is chosen from a set of certain transition matrices defined by a simplex problem, such that the expression $H_v(Q) - \psi_Q(\delta, l)$ is minimized. The value δ is determined by the tolerable error probability. However, gaining $\psi_Q(\delta, l)$ *nats* in the compression is caused by losing of $\Psi_Q(\delta, l)$ amount of information in the transmission, all for the appropriate optimizing Q.

Since Shannon's work in 1948, the problem has been studied by many researchers: Berger, Forney, Blahut, Gray, Ziv, Kieffer, Arimoto, Marton, Dueck, Korner and many others. Our concept based on large deviations theory unifies many previously known techniques within a common framework. This paper presents a consistent and harmonious study of the issue where all the terms, quantities and parameters of the problem conform to create one description of the real problem of finite block-length blockcoding where an error probability is specified and taken into account. Future research might prove successful in obtaining results for classes of sources for which there is a well-developed body of large deviations results for the source output process. In addition, a generalization of the results to multi-parameter sources (random fields) for all D remains an open problem.

Acknowledgement

The author is grateful to Professor Zeev Schuss of Tel-Aviv University for his useful advice.

APPENDIX

An Asymptotic Theory of Large Deviations Theory

Let $\{x_n\}$ be N-dimensional process defined by the stochastic difference equation

$$X_{n+1} = X_n + a_n Z_n \tag{A.1}$$

where $\{z_n\}$ is a N-dim jump process whose conditional jump density at time n is stationary, independent of the values of $z_k, k < n$ and is given by

$$\frac{\partial^N}{\partial z_1 \cdots \partial z_N} \Pr\left\{Z_n \leq z \mid X_n = x, X_{n-1} = x_{n-1} \ldots X_0 = x_0\right\} = w(z, x) \tag{A.2}$$

We assume that the conditional moments of Z_n exist for all $k_1 \ldots k_n$ and are given by

$$m_{k_1 \ldots k_N}(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} z_1^{k_1} \cdots z_N{}^{k_N} w(z_1 \ldots z_N, x)\, \mathrm{d}z_1 \cdots z_N \tag{A.3}$$

Let the sequences $\{a_n\}$ depend upon the time integer n such that

$$\lim_{n \to \infty} a_n = 0 \tag{A.4}$$

$$p(y, n : x, m) = \frac{\partial^N}{\partial y_1 \cdots \partial y_N} \Pr\left\{X_n \leq y | X_m = x\right\} \tag{A.5}$$

for $n > m$. This is the solution, with respect to the "forward" Kolmogorov Equation (or Master Equation) (Gardiner, 1985). We are mostly interested in the case where $a_n = \frac{1}{n+1}$.

$$X_{n+1} = X_n + \frac{1}{n+1} Z_n \tag{A.6}$$

Particularly we consider the case with $Z_n = -bX_n + \xi_n$ where ξ_n is a sequence of zero mean, i.i.d. N-dim random variable, independent of $X_1 \ldots X_n$ with a density function $w(z)$ defined by

$$w(z) = \Pr\left\{\xi_n = z | X_n = x, X_{n-1} = x_{n-1} \ldots X_0 = x_0\right\} \tag{A.7}$$

and conditional moments defined by

$$m_{k_1 \ldots k_N} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} z_1^{k_1} \cdots z_N{}^{k_N} w(z_1 \ldots z_N)\, \mathrm{d}z_1 \cdots z_N \tag{A.8}$$

The most important case related to Source Coding is when $b = 1$.

$$Z_n = -X_n + \xi_n \tag{A.9}$$

We are interested at the evaluation of the probability density function:

$$p(y, n) = \frac{\partial^N}{\partial y_1 \cdots \partial y_N} \Pr\left\{X_n \leq y\right\} \tag{A.10}$$

for $n \gg 1$. Next, the one dimensional theory of (Knessl *et al.*, 1985) is generalized to higher dimensions. The density function $p(y,n;x,m)$ satisfies the forward Kolmogorov equation (Master Equation) (Gardiner, 1985).

$$p(y,n+1;x,m) - p(y,n;x,m)$$

$$= L_y^* p \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big(p(y - \tfrac{z}{n+1}, n; x, m) w(z, y - \tfrac{z}{n+1}) \qquad \text{(A.11)}$$

$$- p(y,n,x,m) w(z,y) \Big) \, \mathrm{d}z_1 \cdots dz_N$$

with the initial condition:

$$p(y,m;x,m) = \delta(y-x) \qquad \text{(A.12)}$$

where $\delta(x)$ is the Dirac measure.

We assume the influences of the initial arguments (x,m) – are suppressed for $n \gg 1$ and taking the case where $Z_n = -bX_n + \xi_n$ and (A.7) we have:

$$\Pr\Big\{Z_n = z | X_n = x\Big\} = \Pr\Big\{-bX_n + \xi_n = z | X_n = x\Big\} = w(z+bx) \qquad \text{(A.13)}$$

$$\Pr\Big\{X_{n+1} \le y\Big\} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \int_{-\infty}^{y_1 - \frac{z_1}{n+1}}$$

$$\cdots \int_{-\infty}^{y_N - z_N/(n+1)} p(X_n = x) p(Z_n = z | X_n = x) \, \mathrm{d}x_1 \cdots \mathrm{d}x_N \, \mathrm{d}z_1 \cdots \mathrm{d}z_N \qquad \text{(A.14)}$$

By differentiation:

$$p(y,n+1) = \frac{\partial^N}{\partial y_1 \cdots \partial y_N} \Pr\Big\{X_{n+1} \le y\Big\}$$

$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[p\Big(y - \tfrac{z}{n+1}, n\Big) w\Big(z + b\Big(y - \tfrac{z}{n+1}\Big)\Big)\Big] \, \mathrm{d}z_1 \cdots \mathrm{d}z_N \qquad \text{(A.15)}$$

Changing variables: $t_i = z_i + by_i - bz_i/(n+1)$ for $i = 1 \ldots N$ and therefore:

$$\Big((n+1)/(n+1-b)\Big) \, \mathrm{d}t_i = \mathrm{d}z_i$$

$$p(y,n+1) = \frac{\partial^N}{\partial y_1 \cdots \partial y_N} \Pr\Big\{X_{n+1} \le y\Big\}$$

$$= \Big(\frac{n+1}{n+1-b}\Big)^N \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[p\Big(y - \tfrac{z}{n+1}, n\Big) w(t)\Big] \, \mathrm{d}t_1 \cdots \mathrm{d}t_N \qquad \text{(A.16)}$$

Using WKB Method (Bender and Orszag, 1987) we construct an approximation solution of (A.16) for $n \gg 1$.

$$p(y,n) \sim \sqrt{n}^N \exp\Big(-n\Psi(y)\Big)\Big(K_0(y) + K_1(y)/n + \cdots\Big) \tag{A.17}$$

The leading term $\Psi(y)$ is determined by substituting $(A.17)$ into $(A.16)$, expanding for large n and equating the coefficients of each power of n to zero. The leading order equation is:

$$\Psi(y) = by\nabla\Psi(y) - \log M\Big(\nabla\Psi(y)\Big) \tag{A.18}$$

where $M(\theta)$ is the moment generating function of ξ_n defined by:

$$M(\theta) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[\exp(\theta z) w(z)\Big]\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \tag{A.19}$$

We substitute the results and have:

$$\begin{aligned}
p(y,n+1) &= \frac{\partial^N}{\partial y_1 \cdots \partial y_N} \Pr\Big\{X_{n+1} \le y\Big\} \\
&= \Big(\frac{n+1}{n+1-b}\Big)^N \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[p\Big(y + \frac{by-t}{1+n-b}, n\Big) w(t)\Big]\, \mathrm{d}t_1 \cdots \mathrm{d}t_N \\
&= \Big(\sqrt{n}\frac{n+1}{n+1-b}\Big)^N \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[\exp\Big(-n\Psi\Big(y + \frac{by-t}{1+n-b}\Big)\Big)\Big] w(t) \\
&\times \Big(K_0(y) + \frac{K_1(y)}{n} + \cdots\Big)\, \mathrm{d}t_1 \cdots \mathrm{d}t_N \\
&= \sqrt{n+1}^N \exp\Big(-(n+1)\Psi(y)\Big)\Big(K_0(y) + \frac{K_1(y)}{n} + \cdots\Big)
\end{aligned} \tag{A.20}$$

Taking the approximation:

$$n\Psi\Big(y + \frac{by-t}{1+n-b}\Big) \sim n\Psi(y) + (by-t)\nabla\Psi(y) \tag{A.21}$$

and comparing $\mathcal{O}(1)$ terms:

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \Big[\exp\Big(\Psi(y) + (by-t)\nabla\Psi(y)\Big)\Big] w(t)\, \mathrm{d}t_1 \cdots \mathrm{d}t_N = 1 \tag{A.22}$$

$$\Psi(y) = by\nabla\Psi(y) - \log M\Big(\nabla\Psi(y)\Big) \tag{A.23}$$

We are interested in Cramer Problem where $b = 1$:

$$\Psi(y) = y\nabla\Psi(y) - \log M\Big(\nabla\Psi(y)\Big) \tag{A.24}$$

Which is equivalent to:

$$\Psi(y) = \sup_{\theta}\Big(\theta y - \log M(\theta)\Big) \tag{A.25}$$

Since:

$$\bar{y}_j = \frac{1}{M(\nabla\Psi)}\left(\frac{\partial(M(\nabla\Psi))}{\partial\Big(\frac{\partial\Psi}{\partial y_j}\Big)}\right) \tag{A.26}$$

is a condition for maximum of: $\Big(\theta y - \log M(\theta)\Big)$. To compute the function $p(y,n)$ we similarly compute $K_0(y), K_1(y)$, by comparing the next coefficient of powers of n at (A.17). It is sufficient to compute $K_0(y)$ for almost all practical applications. We shall collect all $\mathcal{O}(1/n)$ terms in (A.17). Rewrite (A.15):

$$\begin{aligned} p(y,n+1) &= \frac{\partial^N}{\partial y_1\cdots\partial y_N}\Pr\Big\{X_{n+1}\le y\Big\} \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\Big[p\Big(y-\tfrac{z}{n+1},n\Big)w\Big(z+b\Big(y-\tfrac{z}{n+1}\Big)\Big)\Big]\,\mathrm{d}z_1\cdots\mathrm{d}z_N \end{aligned} \tag{A.27}$$

and

$$\begin{aligned} p\Big(\bar{y}-\tfrac{\bar{z}}{n+1},n\Big) \sim \sqrt{n}^N \exp\Big(-n\Big(\Psi(y) - \tfrac{z}{n+1}\nabla\Psi(y)^T\Big) \\ +\tfrac{1}{2(n+1)^2}z\Big[\tfrac{\partial^2\Psi}{\partial y_i\partial y_j}\Big]\bar{z}^T\Big) \\ \times\left(K_0(y) + \frac{K_1(y-z/(n+1))}{n} - \tfrac{z}{n+1}\nabla K_0(y)^T\right) \end{aligned} \tag{A.28}$$

$$\begin{aligned} p\Big(\bar{y}-\tfrac{\bar{z}}{n+1},n\Big) \sim \sqrt{n}^N \exp\Big(-n\Psi(y) + \bar{z}\nabla\Psi(y)^T - \tfrac{1}{n+1}\bar{z}\nabla\Psi(y)^T \\ -\tfrac{1}{2(n+1)}\bar{z}\Big[\tfrac{\partial^2\Psi}{\partial y_i\partial y_j}\Big]\bar{z}^T\Big) \\ \times\left(K_0(y) + \frac{K_1(y-z/(n+1))}{n} - \tfrac{z}{n+1}\nabla K_0(y)^T\right) \end{aligned} \tag{A.29}$$

and

$$w\Big(\bar{z}+b(y-z/(n+1)\Big) = w(\bar{z}+b\bar{y}) - b/(n+1)\bar{z}\nabla w(\bar{z}+b\bar{y}) \tag{A.30}$$

$$p(y,n+1) \sim \sqrt{n+1}^N \exp\Big(-(n+1)\Psi(y)\Big)\left(K_0(y) + \frac{K_1(y)}{n+1} + \cdots\right) \tag{A.31}$$

We use the approximation,

$$\left(\frac{n}{n+1}\right)^{N/2} \sim 1 - \frac{N}{2(n+1)} \tag{A.32}$$

By substituting (A.28)–(A.32) into (A.27) and equating the coefficients of $\mathcal{O}(1/n^2)$ to zero:

$$\begin{aligned} 0 = &\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big) w(\bar{z} + b\bar{y}) \bar{z}\nabla^T K_0(y)\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \\ &+ K_0(y)\Bigg(N/2 + \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} b\bar{z}\nabla w(\bar{z} + b\bar{y}) \\ &\times \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big)\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \\ &+ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left(\bar{z}\nabla\Psi(y)^T + \frac{1}{2}\bar{z}\left[\frac{\partial^2\Psi}{\partial y_i \partial y_j}\right]\bar{z}^T\right) \\ &\times \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big) w(\bar{z} + b\bar{y})\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \Bigg) \end{aligned} \tag{A.33}$$

Computing separately the first term:

$$\begin{aligned} &\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big) w(\bar{z} + b\bar{y}) \bar{z}\nabla^T K_0(y)\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \\ &= \left(b\bar{y} - \frac{1}{M(\nabla\Psi)}\nabla_{\nabla\Psi} M(\nabla\Psi)\right)\nabla^T K_0(y) \\ &= (b-1)\nabla\Psi(y)\left[\frac{\partial^2\Psi}{\partial y_i \partial y_j}\right]^{-1}\nabla^T K_0(y) \end{aligned} \tag{A.34}$$

and we obtain $N-$ dimensional "transport" equation for $K_0(y)$:

$$\begin{aligned} &(1-b)\nabla\Psi(y)\left[\frac{\partial^2\Psi}{\partial y_i \partial y_j}\right]^{-1}\nabla^T K_0(y) - K_0(y)\Bigg(N/2 + \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} b\bar{z}\nabla w(\bar{z} + b\bar{y}) \\ &\times \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big)\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \\ &+ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left(\bar{z}\nabla\Psi(y)^T + \frac{1}{2}\bar{z}\left[\frac{\partial^2\Psi}{\partial y_i \partial y_j}\right]\bar{z}^T\right) \\ &\times \exp\Big(\Psi(y) + \bar{z}\nabla\Psi(y)^T\Big) w(\bar{z} + b\bar{y})\, \mathrm{d}z_1 \cdots \mathrm{d}z_N \Bigg) = 0 \end{aligned} \tag{A.35}$$

We define the matrices C, D as:

$$C_{i,j} = \frac{\partial^2\Psi}{\partial y_i \partial y_j}; \qquad E_{i,j,k} = \frac{\partial^3\Psi}{\partial y_i \partial y_j \partial y_k} \tag{A.36}$$

and by tensor rules:

$$(1-b)\frac{\partial\Psi}{\partial y_i}=\sum_{j=1}^{N}\left(by_j-\frac{1}{M}\frac{\partial M}{\partial(\partial_j\Psi)}\right)\frac{\partial^2\Psi}{\partial y_i\partial y_j} \tag{A.37}$$

and second derivation:

$$\begin{aligned}(1-b)\frac{\partial^2\Psi}{\partial y_i\partial y_k}&=b\frac{\partial^2\Psi}{\partial y_i\partial y_j}+b\sum_{j=1}^{N}\left(y_j\frac{\partial^3\Psi}{\partial y_i\partial y_j\partial y_k}\right)-\frac{1}{M}\sum_{j=1}^{N}\left(\frac{\partial M}{\partial(\partial_j\Psi)}\right)\frac{\partial^3\Psi}{\partial y_i\partial y_j\partial y_k}\\&+\frac{1}{M^2}\sum_{l=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial M}{\partial(\partial_j\Psi)}\right)\frac{\partial^2\Psi}{\partial y_i\partial y_j}\left(\frac{\partial M}{\partial(\partial_l\Psi)}\right)\frac{\partial^2\Psi}{\partial y_l\partial y_k}\\&-\frac{1}{M}\sum_{l=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial^2 M}{\partial(\partial_j\Psi)\partial(\partial_l\Psi)}\right)\frac{\partial^2\Psi}{\partial y_i\partial y_j}\frac{\partial^2\Psi}{\partial y_l\partial y_k}\end{aligned} \tag{A.38}$$

For convenience we denote:

$$\Psi_{ijk}=\frac{\partial^3\Psi}{\partial y_i\partial y_j\partial y_k} \tag{A.39}$$

and $\bar{z}^i$ as the i-th component of vector $\bar{z}$. The derivatives are all under summation – unless specified.

$$\partial_i\ln K_0=\frac{\partial K_0}{\partial y_i}1/K_0 \tag{A.40}$$

Substituting (A.40), (A.37) and the following identities into (A.34):

$$\begin{aligned}&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\exp\left(\Psi(\bar{y})+\bar{z}\nabla\Psi(\bar{y})^T\right)w(\bar{z}+b\bar{y})\bar{z}\nabla^T K_0(y)\,\mathrm{d}z_1\cdots\mathrm{d}z_N\\&\quad=\sum_{i=1}^{N}\left(\frac{1}{M}\frac{\partial M}{\partial(\partial_i\Psi)}-b\bar{y}_i\right)\partial_i K_0\end{aligned} \tag{A.41}$$

$$\left(\partial_i K_0\partial_k\Psi\right)C_{ik}^{-1}=\frac{1}{2}C_{ik}^{-1}K_0\left(\partial_l\Psi C_{jl}^{-1}E_{ijk}+(b-1)\partial_i\Psi\partial_k\Psi\right) \tag{A.42}$$

$$\begin{aligned}&\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\partial^2\Psi}{\partial y_i\partial y_k}^{-1}\right)\frac{\partial\Psi}{\partial y_l}\left(\frac{\partial^2\Psi}{\partial y_j\partial y_l}^{-1}\right)\frac{\partial^3\Psi}{\partial y_i\partial y_k\partial_j}\\&\quad=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\partial^2\Psi}{\partial y_i\partial y_k}^{-1}\right)\frac{\partial\Psi}{\partial y_k}\left(\frac{\partial^2\Psi}{\partial y_j\partial y_l}^{-1}\right)\frac{\partial^3\Psi}{\partial y_i\partial y_l\partial_j}\end{aligned} \tag{A.43}$$

and we obtain:

$$\left(\partial_k\Psi C_{ik}^{-1}\right)\left(\partial_i(\ln K_0)-1/2E_{ijl}C_{jl}^{-1}-1/2(b-1)\partial_i\Psi\right)=0 \tag{A.44}$$

We define:

$$B^i = \left(\partial_i(\ln K_0) - 1/2E_{ijl}C_{jl}^{-1} - 1/2(b-1)\partial_i\Psi\right) \tag{A.45}$$

which satisfies for all i:

$$\partial_k\Psi C_{ik}^{-1}B^i = 0 \tag{A.46}$$

We require $\Psi(\bar{y}) > 0$ for all $\bar{y} \neq 0$.

Rewrite $\Psi(\bar{y})$:

$$\Psi(\bar{y}) = b\bar{y}\nabla\Psi(\bar{y}) - \log M\left(\nabla\Psi(\bar{y})\right) \tag{A.47}$$

derive it,

$$\partial_i\Psi(\bar{y}) = b\partial_i\Psi(\bar{y}) + b\bar{y}_j\partial_{ij}^2\Psi(\bar{y}) - \frac{1}{M\left(\nabla\Psi(\bar{y})\right)}\frac{\partial M}{\partial(\partial_j\Psi)}\partial_{ij}^2\Psi(\bar{y}) \tag{A.48}$$

Therefore $\nabla\Psi(y)$, C are both non-zero for all $\bar{y} \neq 0$. Rewrite (A.45):

$$B^i = \left(\partial_i(\ln K_0) - 1/2(\partial_i C_{jl})C_{jl}^{-1} - \frac{(b-1)\partial_i\Psi}{2}\right) = 0 \tag{A.49}$$

Let M_{jl} denote the determinant of the j, l minor obtained from C.

$$\det C = \sum_{j=1}^{N}(-1)^{j+l}C_{jl}M_{jl} \tag{A.50}$$

$$C_{lj}^{-1} = \frac{Adj[C]_{lj}}{\det C} = \frac{(-1)^{l+j}M_{jl}}{\det C} = \frac{\frac{\partial \det C}{\partial C_{jl}}}{\det C} \tag{A.51}$$

$$\frac{\partial C_{jl}}{\partial y_i}C_{lj} = \frac{\partial \det C}{\partial y_i}\frac{1}{\det C} = \partial_i \ln(\det C) \tag{A.52}$$

Using (A.52), (A.49), (A.46) and we have:

$$\left(\partial_i(\ln K_0) - \frac{1}{2}(\partial_i \ln \det C) - \frac{1}{2}(b-1)\partial_i\Psi\right) = 0 \tag{A.53}$$

$$\frac{\partial\left(\ln\frac{K_0(\bar{y})}{(\det C)^{1/2}} - (b-1)/2\Psi(\bar{y})\right)}{\partial y_i} = 0 \tag{A.54}$$

for $i = 1 \cdots N$. Hence, we have:

$$K_0(\bar{y}) = (2\pi)^{-N/2}\sqrt{\det C}\exp\left\{(b-1)/2\Psi(\bar{y})\right\} \tag{A.55}$$

The higher order terms $K_j(\bar{y})$, $j \geq 1$ can be neglected for our purpose. However, they can be calculated in a similar way. For our purpose $b = 1$. Therefore the approximated density function is,

$$p(\bar{y}, n) = (2\pi/n)^{-N/2}\sqrt{\det C}\exp\left\{-n\Psi(\bar{y})\right\} \tag{A.56}$$

where:

$$\Psi(\bar{y}) = \sup_{\bar{\theta}}\left(\bar{\theta}\bar{y} - \log M(\bar{\theta})\right) \tag{A.57}$$

$$C_{i,j} = \frac{\partial^2 \Psi}{\partial y_i \partial y_j} \tag{A.58}$$

$$M(\bar{\theta}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left[\exp(\bar{\theta}\bar{z})w(\bar{z})\right] \mathrm{d}z_1 \cdots \mathrm{d}z_N \tag{A.59}$$

and

$$w(\bar{z}) = \Pr\left\{\bar{\xi}_n = \bar{z} | \bar{X}_n = \bar{x}, \bar{X}_{n-1} = \bar{x}_{n-1} \ldots \bar{X}_0 = \bar{x}_0\right\} \tag{A.60}$$

References

Algoet P.H. and Cover T.M. (1988): *A sandwich proof of the Shannon-McMillan-Breiman theorem.* — The Annals of Probability, v.16, pp.899–909.

Anantharam V. (1990): *A large deviations approach to error exponents in source coding and hypothesis testing* . — IEEE Trans. Inform. Theory, v.IT–36, pp.938–943.

Arimoto S. (1973): *On the converse to the coding theorem for discrete memoryless channels.* — IEEE Trans. Inform. Theory, v.IT–19, pp.357–359.

Barron A.R. (1985): *The strong ergodic theorem for densities generalized Shannon-McMillan-Breiman theorem.* — Ann. Probability, v.13, pp.1292–1303.

Bender C.M. and Orszag S.A. (1987): *Advanced Mathematical Methods for Scientists and Engineers.* — Englewood Cliffs: McGraw-Hill.

Berger T. (1971): *Rate Distortion Theory: A Mathematical Basis for Data Compression.* — Englewood Cliffs: Prentice-Hall, N.J.

Blahut R.E. (1972): *Computation of channel capacity and rate distortion functions.* — IEEE Trans. Inform. Theory, v.IT–18, pp.460–473.

Blahut R.E. (1974): *Hypothesis testing and information theory.* — IEEE Trans. Inform. Theory, v.IT–20, pp.405–417.

Blahut R.E. (1976): *Information bounds of the Fano-Kullback type.* — IEEE Trans. Inform. Theory, v.IT–22, pp.410–421.

Blahut R.E. (1987): *Principles and Practice of Information Theory.* — Reading MA: Addison-Wesley Publishing Co.

Bleistein N. and Handelsman R.A. (1975): *Asymptotic Expansions of Integrals.* — Holt, Rinehart Winston.

Breiman L. (1957): *The individual ergodic theorem of information theory.* — Ann. Math. Stat., v.28 (corrected in v.31, pp.809–810), pp.809–811.

Covo Y. (1992): *Error Bounds for Noiseless Channels by an Asymptotic Large Deviations Theory.* — M.Sc. Thesis. Tel-Aviv University, Applied Mathematics Dept.

Covo Y. and Schuss Z. (1991): *Error bounds for noiseless channels by an asymptotic large deviations theory.* — Preliminary Report, Tel-Aviv University, Applied Mathematics Dept.

Csiszar I. and Longo G. (1971): *On the error exponent for source coding and for testing simple statistical hypothesis.* — Hungarian Academy of Sciences, *Studia Scientiarum Math. Hungarica*, Budapest, v.6, pp.181–191.

Davisson L.D., Longo G. and Sgarro A. (1981): *The error exponent for the noiseless encoding of finite ergodic Markov sources.* — IEEE Trans. Inform. Theory, v.IT–27, pp.431–438.

Dueck G. and Korner J. (1979): *Reliability function of a discrete memoryless channel at rates above capacity.* — IEEE Trans. Inform. Theory, v.IT–25, pp.82–85.

Eyuboglu M.V. and Forney G.D. (1993): *Lattice and trellis quantization with lattice and trellis bounded codebooks – high rate theory for memoryless sources.* — IEEE Trans. Inform. Theory, v.IT–39, pp.46–59.

Gardiner C.W. (1985): *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences.* — Berlin: Springer-Verlag.

Gray R.M. (1975): *Sliding block source coding.* — IEEE Trans. Inform. Theory, v.IT–21, pp.357–368.

Gray R.M. (1990): *Entropy and Information Theory.* — Berlin: Springer-Verlag.

Gray R.M., Neuhoff D.L. and Omura J.K. (1975): *Process definitions of distortion rate functions and source coding theorems.* — IEEE Trans. Inform. Theory, v.IT–21, pp.524–532.

Kieffer J.C. (1978): *A unified approach to weak universal source coding.* — IEEE Trans. Inform. Theory, v.IT–24, No.6, pp.674–682

Kieffer J.C. (1993): *A survey of the theory of source coding with a fidelity criterion.* — IEEE Trans. Inform. Theory, v.IT–39, No.5, pp.1473–1490.

Knessl C., Matkowsky B.J., Schuss Z. and Tier C. (1985): *An asymptotic theory of large deviations for Markov jump processes.* — SIAM J. Appl. Math., v.46, No.6, pp.1006–1028.

Longo G. and Sgarro A. (1979): *The source coding theorem revisited: a combinatorial approach.* — IEEE Trans. Inform. Theory, v.IT–25, pp.544–548.

Mackenthun K.M. and Pursley M.B. (1978): *Variable rate universal block source coding subject to a fidelity constraint.* — IEEE Trans. Inform. Theory, v.IT–24, No.3, pp.340–360.

Marton K. (1974): *Error exponent for source coding with a fidelity criterion.* — IEEE Trans. Inform. Theory, v.IT–20, pp.197–199

Merhav N. and Neuhoff D.L. (1992): *Variable-to-fixed length codes provide better large deviations performance than fixed-to-variable length codes.* — IEEE Trans. Inform. Theory, v.IT–38, pp.135–140.

Natarajan S. (1985): *Large deviations hypothesis testing and source coding for finite Markov chains.* — IEEE Trans. Inform. Theory, v.IT–31, pp.360–365.

Neuhoff D.L. (1975): *Fixed rate universal block source coding with a fidelity criterion.* — IEEE Trans. Inform. Theory, v.IT–21, No.5, pp.511–523.

Omura J. (1973): *A coding theorem for discrete time sources.* — IEEE Trans. Inform. Theory, v.IT–19, pp.490–498.

Orey S. (1985): *On the Shannon-Perez-Moy theorem.* — Contemp. Math., v.41, pp.319–327.

Ornstein D.S. and Shields P.C. (1990): *Universal almost sure data compression.* — The Annals of Probability, v.18, pp.441–452.

Sadowsky and Bucklew (1990): *On large deviation theory and asymptotically efficient Monte Carlo estimating.* — IEEE Trans. Inform. Theory, v.36, No.3, pp.579–589.

Shannon C.E. (1948): *A mathematical theory of communication.* — Bell Systems Tech. J., v.27, pp.379–423, 623–656.

Shannon C.E. (1959): *Coding theorems for a discrete source with a fidelty criterion.* — IRE Nat. Conv. Rec., Part 4, pp.142–163.

Ziv J. (1972): *Coding of sources with unknown statistics* — Part 1: *Probability of encoding error;* Part 2: *Distortion relative to a fidelity criterion.* — IEEE Trans. Inform. Theory, v.IT–18, pp.384–394.

Received: December 27, 1994