Sequential Source Coding: An optimization viewpoint

V.S. Borkar[1], S.K. Mitter[2], Anant Sahai[3], Sekhar Tatikonda[4]
[1]School of Technology and Computer Science
Tata Institute of Fundamental Research, India
[2]Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA
[3]Dept. of Electrical and Computer Engineering
University of California at Berkeley
[4]Department of Electrical Engineering
Yale University

***Abstract*— The problem of sequential source coding is to minimize the average entropy rate subject to a constraint on the average distortion and a causality constraint on codewords. This is cast as an optimization problem on an appropriate convex set of probability measures. Existence and properties of optimal sequential codes are explored. A sequential rate distortion theorem is proved and a construction given to show that in general, a "causality gap" exists.**

I. INTRODUCTION

The technology of distributed and networked control requires understanding the behavior of interconnections of communication and control systems. This requires a new theory of Information which is dynamical and where the constraints of causality and delays are explicitly taken into account (see for example, Mitter, S.K., "Control with Limited Information," Eur. Jrn. Control, Vol. 7, pp. 122-131, 2001.). Rate Distortion Theory, as developed by Shannon, is asymptotic in nature and tells us what the fundamental limitations to reliable transmission subject to distortion constraints are. In earlier work (S. Tatikonda, A. Sahai and S.K. Mitter, "Stochastic Linear Control Over a Communication Channel," IEEE Trans. on Auto Control, Special Issue on Networked Control Systems, Vol. 49, Iss. 9, Sept. 2004, pp. 1549-1561.) the role of sequential rate distortion theory in obtaining lower bounds to performance of LQG with communication constraints has been demonstrated.

The aim of this paper is to cast the sequential source coding problem in its traditional 'rate-distortion' framework as an optimization problem on a convex set of probability measures. This allows us to obtain interesting results about the structure of optimal sequential codes and take the first steps towards a 'sequential rate-distortion theory'. See [11], [12], [15] for related viewpoints. Our work differs from these in its approach and also in that we do not impose any i.i.d. condition on the source as in [11] or stationarity condition on the source-code pair as in the other two (though we do assume the source itself to be stationary). Similar techniques have been applied in [2] to Markov decision processes. Also, see [4] for a related but different perspective of the sequential source coding problem for the special case of Markov sources.

The paper is organized as follows: First, we review the traditional formulation of rate-distortion theory. The next section introduces our formulation of the source-code pair as a probability measure on a product space. The latter is a product of the countable product of the source alphabet and the countable product of the code alphabet. A probability measure on this space then corresponds to a canonical realization of the joint source-code process. Since the source is fixed, the marginal of this measure on the first factor space (i.e., countable product of source alphabet) gets fixed. The regular conditional law on the second factor space (i.e., countable product of the code alphabet) is then identified with a possibly randomized encoding of the source. This permits us to view the family of source-code pairs for a fixed source as a family of probability measures on the above product space with a fixed marginal. This set is closed and convex with its extreme points corresponding to deterministic codes wherein the aforementioned regular conditional law is a.s. a Dirac measure. We shall be interested in the closed convex subset corresponding to sequential or 'causal' codes which satisfy an additional causality constraint: an encoding is causal if at each time instant, the encodings up to the time are conditionally independent of the future source outputs given the past. The extremal elements of this set once again correspond to the codes that are sequential and deterministic. Codes that randomize between at most finitely many of these are of special interest to us and are dubbed 'finitely randomized codes'. This section proves various basic properties of these sets of probability measures.

Section III introduces the two optimization problems of concern here, viz., those of minimizing the average information rate of the code and the average mutual information of the source-code pair, both subject to a given bound on the average distortion. While the former is the usual rate-distortion problem, the latter is motivated purely by analogy with Shannon's rate distortion theorem which identifies the two problems in the block coding paradigm. For a finite time horizon, these problems amount to minimizing a concave (resp. convex) function on a convex set. Using some facts

from convex analysis, the minimum in the former problem can be shown to be attained by a finitely randomized code. For infinite time horizon, the limiting time average may not be defined and one must consider both the $\limsup$ and the $\liminf$ in either case. Further, we can also consider the limit of the minima of finite horizon problems as the horizon extends to infinity. The main result of this section is that for either problem, each formulation leads to the same number.

Section III shows that the equality in fact extends across the problems if one is willing to restrict to finitely randomized codes. That is, the minimum of $\limsup/\liminf$ and the limit of minima for finite horizon problems for rate minimization are all equal to each other and to the minima of $\limsup/\liminf$ for the mutual information minimization problem *if* one restricts to finitely randomized codes in the latter case. This captures in some sense the spirit of Shannon's rate distortion theorem for the sequential case, because of which we call it the 'sequential rate distortion theorem'. The minima of $\limsup/\liminf$ and the limit of minima of finite horizon problems in the minimization of mutual information that allows for arbitrary randomizations are equal and upper bounded by the minimum of the corresponding entropy minimization problem mentioned above. This inequality can be strict (leading to a "causality gap"), as shown by a constructive counterexample in Section V. Section VI concludes with some related remarks.

We shall now summarize our main results (Theorem 3.1–Corollary 3.4 below). We consider a stationary source $\{X_n\}$ taking values in a finite alphabet and an associated process of its encodings $\{Z_n\}$ such that Z_n encodes X_n, based on *only* the 'history' $X_m, Z_m, m < n$, and possibly an additional randomization device that does not, however, anticipate the future outputs of the source. In other words, we shall be looking at the probability measures induced by the source-code pair on the product path space, with the 'marginal' corresponding to the source being a fixed stationary measure. Such a point of view is not entirely new, being implicit in the works of Gray et al. [10], [11], [12], *sans* the sequentiality hypothesis. We shall be interested in particular in the special subclass corresponding to finitely randomized sequential codes, i.e., the codes obtained by randomizing between finitely many deterministic sequential codes. It is to be kept in mind that this randomization is *a priori*, i.e., at time zero, the code once picked thus is deterministic thereafter.

It is worth noting that in contrast to Shannon theory where such randomization, if and when used, is purely a technical device, it does make a nontrivial difference in the sequential case, as we see below.

We shall be interested in the following quantities :

(i) infimum over all sequential codes subject to the distortion constraint, of the $\liminf$ of the time-averaged entropy (entropy rate) of codewords,

(ii) same as above, with $\limsup$ in place of $\liminf$,

(iii) same as (i), with infimum over finitely randomized sequential codes,

(iv) same as (ii), with infimum over finitely randomized codes,

(v) limit (shown to exist) as the time horizon recedes to infinity, of the infimum over all sequential codes (equivalently, all finitely randomized sequential codes) of the per letter entropy over a finite horizon, subject to the distortion constraint,

(vi) same as *(i)*, but with source-code mutual information (s.c.m.i.) in place of entropy,

(vii) same as *(ii)*, with s.c.m.i. in place of entropy,

(viii) same as *(iii)*, with s.c.m.i. in place of entropy,

(ix) same as *(iv)*, but with s.c.m.i. in place of entropy,

(x) same as *(v)*, but with s.c.m.i. in place of entropy.

Our main result is that *(i)–(v)* and *(viii)–(ix)* agree with each other, so do *(vi)*, *(vii)*, *(x)* and the latter are a lower bound for the former.

The equality of *(i)–(v)*, *(viii)–(ix)* is reminiscent of Shannon's rate distortion theorem. Since finite randomizations amount to taking convex combinations in the space of probability measures, what we have here is a statement that the 'rate' function is precisely what one obtains by taking source-code mutual information for finite convex combinations of deterministic codes, followed by the appropriate limits / infimum.

At the same time, the gap between the equal quantities *(i)–(v)* and *(viii)–(ix)* on one hand and *(vi)–(vii)*, *(x)* on the other can be a strict gap, as we show by a concrete example. This shows conclusively that a 'full' rate distortion theorem as in the block coding case (extended to a very general class of sources in [13], see also [14] for some related results for channel coding) is not possible. Of course, with the additional constraint of sequentiality, one would expect to lose some aspects of the Shannon theory for block coding. Our main result together with our counterexample underscores precisely what this loss is.

While the main aim of this article is to take some initial steps towards a sequential rate distortion theory, going a little beyond [15], it is also hoped that the novel 'optimization' formulation and the added insight it brings are of interest in themselves.

A. Review of Traditional Rate Distortion Theory

Traditional Rate-Distortion theory is a well-established area of information theory dealing with the lossy compression of data from random sources. By "lossy" we mean that we allow the reconstructed data to differ from the original. Once we allow differences, it is natural to want some way of representing the fidelity of the reconstruction between a source sequence x_1^n and its reconstruction/encoding $\hat{x}_1^n$.

Definition 1.1: A *distortion measure* is a family of functions $\rho_n(x_1^n, \hat{x}_1^n)$ which all take values in the positive real numbers. The family is called a *per-letter distortion measure* if there exists a function $\rho(x, \hat{x})$ called the *single-letter distortion function* such that $\rho_n(x_1^n, \hat{x}_1^n) = \frac{1}{n}\sum_{k=1}^{n} \rho(x_k, \hat{x}_k)$ for every positive value of n.

We restrict our attention to per-letter distortion measures and are concerned with the average per-letter distortion

$\frac{1}{n}E[\rho_n(X_1^n, \hat{X}_1^n)]$ and its asymptotic properties. By "compression" of the data, we want to capture the idea of size. The traditional way to do that is to consider the sequence to be generated by a random process and to evaluate the *entropy* of the sequence of random variables. We restrict ourselves to *finite alphabet* random sources which produce symbols from a finite set S.

Definition 1.2: The *entropy* of a sequence of random variables X_1^n is denoted by $H(X_1^n)$. For a single random variable X, conditioned on another variable Y, we represent the *conditional entropy* by $H(X|Y) = \sum_{(x,y)} P(X = x, Y = y)\log(P(X = x|Y = y))$ where $P(X = x, Y = y)$ is the probability of the event $X = x, Y = y$ and $P(X = x|Y = y)$ is the probability of the event $X = x$ conditioned on the fact that the event $Y = y$ has occurred. It turns out that $H(X_1^n) = \sum_{k=1}^n H(X_k|X_{k-1}, X_{k-2}, \cdots)$ (where we use the convention that $X_i =$ an arbitrary fixed element of S for negative i). The *entropy rate* $H(X)$ of a random process is defined to be $\lim_{n\to\infty} \frac{1}{n}H(X_1^n)$ if the limit exists.

All logarithms in this paper are taken in base 2. Entropy is motivated as a measure of size by the following lossless coding theorem. $\{0,1\}^*$ is the set of all finite binary strings.

Theorem 1.1: [6] There exists a sequence of functions (called deterministic encoders) F_n which map $x_1^n \in S^n$ into variable length binary strings from $\{0,1\}^*$ such that

$$\lim_{n\to\infty} \frac{1}{n}E(\text{length}(F_n(X_1^n))) = H(X)$$

and furthermore there exists a sequence of inverse functions (called deterministic decoders) G_n so that $G_n(F_n(x_1^n)) = x_1^n$ for all strings $x_1^n \in S^n$.

Traditional rate-distortion theory often focuses on *memoryless* sources: those for which the $\{X_k\}$ are independent and identically distributed. Then, using Huffman codes, one can achieve within one bit of the average entropy and thus the per letter expected length of the encoding is in fact within $\frac{1}{n}$ of the entropy. (Recall that Huffman codes are uniquely decodable prefix codes.) Furthermore, $H(X) = H(X_k)\ \forall k$ since conditioning is irrelevant.

Given the above definitions and motivations, we want to be able to characterize the minimal entropy rate $H(\hat{X})$ required to achieve a certain average distortion between X and $\hat{X}$.

Definition 1.3: A *deterministic source encoder* of block-length N is a function F_N from source sequences $x_1^N \in S^N$ into reconstruction sequences $\hat{x}_1^N \in \Sigma^N$. F_N, a deterministic source encoder of block-length N, is considered to be *memoryless* if there exists an F_1, deterministic source encoder of block-length 1, such that $F_N(x_1^N) = [F_1(x_1), F_2(x_2), \cdots, F_1(x_N)]$.

Theorem 1.2: [6] For a memoryless random source $\{X_k\}$, define

$$R(K) = \inf_{P(\hat{X}=\hat{x}|X=x)|\sum_{x,\hat{x}} P(X=x)P(\hat{X}=\hat{x}|X=x)\rho(x,\hat{x})\le K} I(X,\hat{X})$$

where $I(X,\hat{X}) = H(\hat{X}) - H(\hat{X}|X)$ is the mutual information between X and $\hat{X}$ if they were related by the joint distribution that had the specified marginal $P(X = x)$ and conditional $P(\hat{X} = \hat{x}|X = x)$. Then, $\forall K, \epsilon$ we have $\exists N$ and $\exists$ deterministic source encoder F_N such that the average distortion $\frac{1}{N}E(\rho_N(X_1^N, F_N(X_1^N))) \le K$ and the average output entropy $\frac{1}{N}H(F_N(X_1^N)) \le R(K)+\epsilon$. Furthermore, no encoders exist which have average output entropy less than $R(K)$ while still having average distortion less than K. In addition we can restrict ourselves to fixed-length codewords.

Notice that the encoders used here are block encoders and are thus not generally causal. That is, for block encoders $\hat{X}^k$ can depend on X^{k+m} where $m > 0$. This is not always an issue since in applications like lossy image compression, when the samples are distributed in space and so causality is not important. However, in the cases where the $\{X_k\}$ represent samples of a random process evolving in time, the non-causality of the encoders in the above theorem can be a problem, especially if the underlying application has a "real-time" flavor to it. For these, we require a general definition of causal encodings:

Definition 1.4: A *causal deterministic source encoder* F is an infinite sequence of functions $[f_1, f_2, \cdots]$ such that f_k maps source sequences $x_1^k \in S^k$ into a single symbol $\hat{x}_k \in \Sigma$ which is the reconstruction at time k. A *memoryless encoder* is one for which $f_k(x_1^k) = f_k(x_k)$ meaning that the encoding of a particular source symbol does not depend on previous symbols.

Definition 1.5: The *rate-distortion performance* of a source encoder F (mapping source sequences x_1^∞ into $\hat{x}_1^\infty$) is the pair (R_F, K_F) where $R_F = \lim_{n\to\infty} \frac{1}{n}H(\hat{X}_1^n)$ is the average output entropy rate and $K_F = \lim_{n\to\infty} \frac{1}{n}E(\rho_n(X_1^n, \hat{X}_1^n))$ is the average distortion.

For memoryless sources and causal deterministic source encoders, the performance region is completely characterized by a general result due to Gilbert and Neuhoff.

Theorem 1.3: ([15], Theorem 3) For the encoding of memoryless sources, causal deterministic source encoders have an operational rate-distortion performance curve which is the lower convex envelope of the operational rate-distortion performance of memoryless encoders. Moreover one can achieve any point on this performance curve by time-sharing between two memoryless encoders.

In [15], the above result is proved for deterministic encoders, which are of primary interest in the real world. However, given that the spirit of source encoding resembles constrained optimization problems, where the usefulness of "mixed strategies" and randomization [2] is well established, we consider in this paper the generalization of source encoders to include randomized encoders. First, we need to carefully define randomized sequential codes and introduce the idea of *finite randomization*.

II. The Structure of Sequential Codes

Our source will be a stationary stochastic process $X_n, n \ge 1$, taking values in a finite alphabet S. (The possibility of allowing more general S, which is indeed possible for most of our results, will be briefly commented upon in the discussion at the end.) At each time n, we have an encoding

of X_n into a Σ-valued random variable Z_n, Σ being another prescribed finite alphabet. (Since S and Σ are finite, they are trivially compact as well.) For $N = 1, 2, \ldots, \infty$, denote by S^N, Σ^N the N-fold product of S, Σ resp. with the product topology and the associated product Borel σ-field. Given an infinite string $x^\infty \stackrel{\Delta}{=} [x_1, x_2, x_3, \ldots,]$, we shall denote by x^n the vector $[x_1, \ldots, x_n]$ and by $\tilde{x}^n$ the string $[x_{n+1}, x_{n+2}, \ldots]$ for $n \geq 0$. Finally, for a Polish space $\chi, \mathcal{P}(\chi)$ will denote the Polish space of probability measures on χ with the Prohorov topology ([3], Chapter 2), and the notation $\mathcal{L}(\ldots)$ will stand for 'the law of $\ldots$'.

For N as above, we specify the source X^N by specifying $\mathcal{L}(X^N) = \nu^N \in \mathcal{P}(S^N)$. Define $\mathcal{D} \subset \mathcal{P}(S^N \times \Sigma^N)$ by

$$\mathcal{D} = \{\mu^N \in \mathcal{P}(S^N \times \Sigma^N) : \mu^N(dx, \Sigma^N) = \nu^N(dx)\},$$

i.e., $\mathcal{D}$ is the set of probability measures on the product space $S^N \times \Sigma^N$ whose marginal on the factor space S^N coincides with the prescribed ν^N. In particular, for X^N, Z^N as above, $\mathcal{L}(X^N, Z^N) \in \mathcal{D}$. Conversely, each $\mu^N \in \mathcal{D}$ can be disintegrated as

$$\mu^N(dx, dy) = \nu^N(dx) q^N(x, dy),$$

where the map $x \in S^N \rightarrow q^N(x, dy) \in \mathcal{P}(\Sigma^N)$ is the regular conditional law specified ν^N-a.s. uniquely ([3], pp.41). We can identify this map with a possibly randomized encoding scheme that generates Z^N given X^N. Thus we have a one-one correspondence between the elements of $\mathcal{D}$ and possibly randomized encodings of X^N. For most of what follows, we have $N = \infty$.

Lemma 2.1: $\mathcal{D}$ is compact convex in $\mathcal{P}(S^N \times \Sigma^N)$.

We now introduce some distinguished subsets of $\mathcal{D}$.

Definition 2.1: The encoding scheme $q^N(\cdot, \cdot)$ is said to be *deterministic* if $q^N(x, dy)$ is a Dirac measure (i.e., support$(q^N(x, dy))$ is a singleton) for ν^N-a.s. x.

Definition 2.2: Call the encoding scheme *sequential* if for each $n \geq 0$, the corresponding Z^n is conditionally independent of future source outputs $\tilde{X}^n$ given the past source outputs X^n. $\mathcal{D}_{seq} \subset \mathcal{D}$ will denote the subset of $\mathcal{D}$ corresponding to sequential codes and $\mathcal{D}'_{seq}$ its further subset that corresponds to deterministic sequential codes.

It is worth noting that this specification does not characterize the joint law of (X^N, Z^N).

Lemma 2.2: $\mathcal{D}_{seq}$ is convex compact.

For $n \geq 1$, let $\Gamma_n = \mathcal{L}(X^n, Z^{n-1}), \Phi_n = \mathcal{L}(X^n, Z^n), x^n \rightarrow p_n(x^n, dx)$ the regular conditional law of X_{n+1} given X^n (prescribed ν^n-a.s. uniquely) and $(x^n, z^{n-1}) \rightarrow \psi_n(dz/x^n, z^{n-1})$ the regular conditional law of Z_n given (X^n, Z^{n-1}) for $n \geq 1$ (prescribed Γ_n-a.s. uniquely). Here $Z^0 =$ an arbitrary fixed element of Σ by convention. Then Φ_n is defined recursively by

$$\Phi_{n+1}(dx^{n+1}, dz^{n+1}) = \Phi_n(dx^n, dz^n) p_n(x^n, dx_{n+1}) \psi_{n+1}(dz_{n+1}/x^{n+1}, z^n), \quad (1)$$

by virtue of Definition 2.2.

We now digress briefly to recall a key result from [1]. Let S_1, S_2 be Polish spaces equipped with their Borel σ-fields and $Q \subset \mathcal{P}(S_1 \times S_2)$ the set of probability measures μ such that $\mu(dx, S_2) = \nu(dx)$ for a prescribed $\nu(dx) \in \mathcal{P}(S_1)$. Clearly, Q is closed convex. Let $Q_e = \{$extreme points of $Q\}$ and $Q_D = \{\mu \in Q : \mu(dx, dy) = \nu(dx) v(x, dy)$ where $v(x, dy)$ is a Dirac measure for ν-a.s. $x\}$.

Lemma 2.3: [1] $Q_e = Q_D$.

Corollary 2.1: The set of extreme points of $\mathcal{D}_{seq}$ coincides with $\mathcal{D}'_{seq}$.

Lemma 2.4: Each $\mu \in \mathcal{D}_{seq}$ is the barycenter of a $\xi \in \mathcal{P}(\mathcal{D}'_{seq})$, i.e., $\int f(\mu') d\xi(\mu') = f(\mu)$ for all affine $f : \mathcal{D}_{seq} \rightarrow \mathcal{R}$.

Remark: That $\mathcal{D}'_{seq}$ is measurable (in fact, G_δ, i.e., a countable intersection of open sets) follows from Corollary 2.1 above and Corollary 27.3, p. 138, [5].

Definition 2.3: A $\mu \in \mathcal{D}_{seq}$ will be said to be *asymptotically deterministic* if the corresponding $\psi_n(dz/X^n, Z^{n-1})$ a.s. converges to a possibly random Dirac measure on Σ as $n \rightarrow \infty$.

Let $\mu \in \mathcal{D}_{seq}$ and $\xi \in \mathcal{P}(\mathcal{D}'_{seq})$ as above and $\mathcal{L}(X^\infty, Z^\infty) = \mu$. Then we can view ξ as a 'prior' on the 'parameter space' $\mathcal{D}'_{seq}$ in a Bayesian set-up as follows: Clearly, $\mu(dx, dy) = \int \xi(d\alpha) \alpha(dx, dy)$. Let γ be a $\mathcal{D}'_{seq}$-valued random variable with law $\xi_0 \stackrel{\Delta}{=} \xi$ and for $n \geq 1$, let $\xi_n(d\alpha/X^n, Z^{n-1})$ denote its regular conditional law given the canonically defined (X^n, Z^{n-1}). We view γ as an unknown parameter and its law ξ_0 as a prior on $\mathcal{D}_{seq}$, and ξ^n's as the corresponding posteriors. For $\eta \in \mathcal{D}_{seq}$, let $\psi_n^\eta(dz/x^n, z^{n-1})$ denote the corresponding $\psi_n(dz/x^n, z^{n-1}), n \geq 0$. Then $\psi_n^\alpha(dz/x^n, z^{n-1})$ is Dirac a.s. for $\alpha \in \mathcal{D}'_{seq}$ and

$$\psi_n^\mu(dz/X^n, Z^{n-1}) = \int \xi_n(d\alpha/X^n, Z^{n-1}) \psi_n^\alpha(dz/X^n, Z^{n-1}), n \geq 0.$$

Lemma 2.5:

$$\xi_n(d\alpha/X^n, Z^{n-1}) \rightarrow \delta_\gamma \text{ a.s.},$$

where δ_u is the Dirac measure at u. Equivalently,

$$E[f(\gamma)/X^n, Z^{n-1}] \rightarrow f(\gamma) \text{ a.s., } f \in C_b(\mathcal{D}'_{seq}).$$

Definition 2.4: A $\mu \in \mathcal{D}_{seq}$ is said to be finitely randomized if the corresponding ξ is finitely supported (i.e., the corresponding γ takes only finitely many values). Let $\tilde{D}_{seq}$ denote the subset of $\mathcal{D}_{seq}$ corresponding to finitely randomized μ's.

Corollary 2.2: Every finitely randomized μ is asymptotically deterministic.

This is a straightforward consequence of Lemma 2.5.

Lemma 2.6: $\tilde{\mathcal{D}}_{seq}$ is dense in $\mathcal{D}_{seq}$.

Alternatively, one can use Corollary 2.1 and the Krein-Milman theorem ([5], pp.105).

III. The Optimization Problem

In the traditional rate distortion theory ([6], Chapter 13), one seeks to minimize the average entropy rate subject to an upper bound on average distortion, over the set of all possible encodings. Given our identification of sequential

codes with an appropriate convex set of probability measures, this suggests that we cast the sequential version of the rate-distortion problem as a constrained optimization problem over this set. This is what we do in this section. Shannon's rate distortion theorem ([6], Chapter 13) identifies this constrained minimization of entropy rate with another constrained minimization, viz., that of mutual information between the source outputs and the codewords, in an asymptotic sense. With this in view, we also consider the ancillary problem of minimizing the average mutual information between the source outputs and the codewords subject to the distortion constraint.

We start with some notation. For

$$\mu^{\infty}(dx, dy) = \nu^{\infty}(dx)q^{\infty}(x, dy) \in \mathcal{D}_{seq} \ ,$$

let $\mu^n(dx, dy) = \nu^n(dx)q^n(x, dy)$ denote its restriction to $\mathcal{P}(S^n \times \Sigma^n), n \geq 1$. Let $\mathcal{L}(X^{\infty}, Z^{\infty}) = \mu^{\infty}$. Let $H_n(\mu^n), I_n(\mu^n)$ denote respectively the Shannon entropy of Z^n and the mutual information $I(X^n; Z^n)$. Also, let $\rho : S \times \Sigma \to \mathcal{R}^+$ be a prescribed per symbol distortion measure. Define the average distortion

$$\begin{aligned} D_n(\mu^n) &\stackrel{\Delta}{=} \int \rho^n(x^n, y^n)\mu^n(dx^n, dy^n), \\ D^*(\mu^n) &\stackrel{\Delta}{=} \limsup_{n\to\infty} \frac{D_n(\mu^n)}{n}, \end{aligned}$$

where

$$\rho^n(x^n, y^n) \stackrel{\Delta}{=} \Sigma_{m=0}^{n-1}\rho(x_i, y_i), n \geq 1.$$

Define $F^*(\mu^{\infty}) = \limsup_{n\to\infty} \frac{F_n(\mu^n)}{n}$, $F_*(\mu^{\infty}) = \liminf_{n\to\infty} \frac{F_n(\mu^n)}{n}$ for $F_n(\cdot) = H_n(\cdot), I_n(\cdot)$ or $D_n(\cdot)$. Let $K, 0 < K < \infty$, be a prescribed constant and let $\mathcal{D}_n = \{\mu^n : \mu^{\infty} \in \mathcal{D}_{seq}, D_n(\mu^n) \leq nK\}$, $n \geq 0, \mathcal{D}^* = \{\mu^{\infty} \in \mathcal{D}_{seq} : D^*(\mu^{\infty}) \leq K\}$. Set $R_n = \inf_{\mathcal{D}_n} H_n(\mu^n), n \geq 0$, $R^* = \inf_{\mathcal{D}^*} H^*(\mu), R_* = \inf_{\mathcal{D}^*} H_*(\mu)$, $J_n = \inf_{D_n} I_n(\mu^n), n \geq 0, J^* = \inf_{\mathcal{D}^*} I^*(\mu)$, $J_* = \inf_{\mathcal{D}^*} I_*(\mu)$.

Note that for $n < \infty, H_n(\cdot)$ (respectively, $I_n(\cdot)$) are concave (respectively convex) and continuous on $\mathcal{D}_{seq}$. (See, e.g., [6], p. 31.) Since pointwise minimum and pointwise limits of concave functions are concave, $H_*(\cdot)$ is concave. Likewise, $I^*(\cdot)$ and $D^*(\cdot)$ are convex.

For $n < \infty$, the rate distortion problem is to minimize $H_n(\cdot)$ over $\mathcal{D}_n$. For $n = \infty$, it is to minimize either $R^*(\cdot)$ or $R_*(\cdot)$ on $\mathcal{D}^*$. The ancillary problems are defined accordingly with $I_n(\cdot)$ replacing $H_n(\cdot)$ (respectively, $J^*(\cdot)/J_*(\cdot)$ replacing $R^*(\cdot)/R_*(\cdot)$). We first consider the case of finite n.

Lemma 3.1: $H_n(\cdot), I_n(\cdot)$ attain their minimum values R_n, J_n respectively on $\mathcal{D}_n$. In the former case the minimum is attained at a $\mu \in \tilde{\mathcal{D}}_{seq}$ which randomizes between at most two deterministic codes.

Lemma 3.2: $R_{\infty} = \lim_{n\to\infty} \frac{R_n}{n}$ exists.

Write R_{∞} as $R_{\infty}(K)$ for $n = 0, 1, \cdots, \infty$, to denote explicitly its dependence on K. Then $R_{\infty}(\cdot)$ is a bounded nonincreasing function and therefore can have at most countably many points of discontinuity. Since it is upper semicontinuous and nonincreasing, it is left-continuous.

Lemma 3.3: For all but at most countably many choices of K, $R_{\infty}(K) = R^* = R_*$.

Let 'generic K' stand for 'all but at most countably many values of K'.

Corollary 3.1: For generic K, the following holds: If $\mu^{\infty} \in \mathcal{D}$ satisfies $H^*(\mu^{\infty}) = R^*$, then $H_*(\mu^{\infty}) = R_* = R^* = H^*(\mu^{\infty}) = R_{\infty}$. More generally, given any $\epsilon > 0$, there exists a $\mu^{\infty} \in \mathcal{D}^*$ such that

$$R_{\infty} = R_* = R^* \leq H_*(\mu^{\infty}) \leq H^*(\mu^{\infty}) \leq R_{\infty} + \epsilon.$$

An exactly parallel treatment is possible for the ancillary problem of minimizing the average mutual information subject to the distortion constraint. We state the results in Lemma 3.5 below.

Lemma 3.4: Let random variables Y_1, Y_2, W_1, W_2 satisfy: W_1 is conditionally independent of (Y_2, W_2) given Y_1 and W_2 is conditionally independent of (Y_1, W_1) given Y_2. Then

$$I((Y_1, Y_2);\ (W_1, W_2)) \leq I(Y_1; W_1) + I(Y_2; W_2).$$

Lemma 3.5: $J_{\infty}(= J_{\infty}(K)) \stackrel{\Delta}{=} \inf_n \frac{J_n}{n} = \lim_{n\to\infty} \frac{J_n}{n}$ and for generic $K, J_{\infty} = J^* = J_*$. Furthermore, for generic K the following holds: If $\mu^{\infty} \in \mathcal{D}^*$ satisfies $I^*(\mu^{\infty}) = J^*$, then $I_*(\mu^{\infty}) = J_* = J^* = I^*(\mu^{\infty})$. More generally, given any $\epsilon > 0$, there exists a $\mu^{\infty} \in \mathcal{D}^*$ such that $J_{\infty} = J_* = J^* < I_*(\mu^{\infty}) \leq I^*(\mu^{\infty}) \leq J_{\infty} + \epsilon$.

We can say a little more for this problem when (X^{∞}, Z^{∞}) is jointly stationary. Let θ denote the shift operator on $S^{\infty} \times \Sigma^{\infty}$, mapping a string $[(w_0, w_1, \ldots), (v_0, v_1, \ldots)]$ to $[(w_1, w_2, \ldots), (v_1, v_2, \ldots)]$. Say that $\mu^{\infty} = \mathcal{L}(X^{\infty}, Z^{\infty}) \in \mathcal{D}_{seq}$ is stationary if $\mu^{\infty} = \mu^{\infty} \circ \theta^{-1}$, equivalently, if (X^{∞}, Z^{∞}) is jointly stationary. Let $\hat{\mathcal{D}} = \{\mu^{\infty} \in \mathcal{D}^* : \mu^{\infty}$ is stationary $\}$.

Lemma 3.6:
$J_{\infty} = \inf_{\hat{\mathcal{D}}} \limsup_{n\to\infty} \frac{I_n(\mu^n)}{n} = \inf_{\hat{\mathcal{D}}} \liminf_{n\to\infty} \frac{I_n(\mu^n)}{n}$.

Recall the set $\tilde{\mathcal{D}}_{seq}$ of finitely randomized sequential $\mu \in \mathcal{D}_{seq}$. Let $\tilde{\mathcal{D}}_n$ denote its restriction to $S^n \times \Sigma^n$ intersected with $\mathcal{D}_n$ for $n \geq 1$ and $\tilde{\mathcal{D}}^* = \mathcal{D}^* \cap \tilde{\mathcal{D}}_{seq}$. Set $\tilde{R}_n = \inf_{\tilde{\mathcal{D}}_n} H_n(\mu^n), \tilde{R}^* = \inf_{\tilde{\mathcal{D}}^*} H^*(\mu)$, $\tilde{R}_* = \inf_{\tilde{\mathcal{D}}^*} H_*(\mu)$, $\tilde{J}_n = \inf_{\tilde{\mathcal{D}}_n} I_n(\mu^n), \tilde{J}^* = \inf_{\tilde{\mathcal{D}}^*} I^*(\mu)$, $\tilde{J}_* = \inf_{\tilde{\mathcal{D}}^*} I_*(\mu)$.

Lemma 3.7:
$\lim_{n\to\infty} \frac{\tilde{R}_n}{n} = \inf_n \frac{\tilde{R}_n}{n} = R_{\infty}, \lim_{n\to\infty} \frac{\tilde{J}_n}{n} = \inf_n \frac{\tilde{J}_n}{n} = J_{\infty}$.

Lemma 3.8: For generic K, $\tilde{R}^*(K) = \tilde{R}_*(K) = R_{\infty}(K)$.

IV. A 'Sequential Rate Distortion Theorem'

We are now ready to establish a result that may be viewed as a sequential version of Shannon's rate distortion theorem. Let $\tilde{\mathcal{D}}^*(m) = \{\mu \in \tilde{\mathcal{D}}^* : |\text{ support } (\mu)| \leq m\}, m \geq 1$, where $|A|$ denotes the cardinality of a set A.

Lemma 4.1: Let $\mu^{\infty} = \mathcal{L}(X^{\infty}, Z^{\infty}) \in \tilde{\mathcal{D}}^*(m), m \leq 1$. Then

$$\frac{H(Z^n/X^n)}{n} \to 0.$$

Corollary 4.1: For
$m \geq 1, \inf_{\tilde{\mathcal{D}}^*(m)} H^*(\mu) = \inf_{\tilde{\mathcal{D}}^*(m)} I^*(\mu)$.

Corollary 4.2: $\tilde{J}^* = \tilde{R}^*$, $\tilde{J}_* = \tilde{R}_*$.

Thus we have :

Theorem 4.1: For generic K, $R_\infty = R^* = R_* = \tilde{R}^* = \tilde{R}_* = \tilde{J}^* = \tilde{J}_* \geq J_\infty = J^* = J_*$.

Let $F(K)$ (resp., $G(K)$) denote any of the equal quantities on the left (resp., right) hand side of the above inequality, with the dependence on K made explicit. The next result shows that the qualification 'for generic K' can be dropped for most purposes.

Corollary 4.3: $F(\cdot)$ (resp., $G(\cdot)$) is convex and therefore continuous on the interior of its domain.

V. A Counterexample

This section shows by a constructive counterexample that the inequality in Theorem 4.1 cannot be replaced by an equality. First, we will show that for i.i.d. sources, any finitely randomized sequential code has both average distortion and entropy rate equal to an appropriate deterministic code to be constructed as follows. There is no point in considering infinitely randomized codes since we have already shown that the minimum output entropy is attained on finitely randomized ones.

Consider a finitely randomized code that randomizes between l deterministic sequential codes $f_k^i : X_1^k \to \hat{X}_k$ where $1 \leq i \leq l$ and $k \geq 1$. The randomization is done with probabilities $P = [p_1, \cdots, p_l]$. Let $(R_i, K_i), 1 \leq i \leq l$ denote the rate-distortion performance pairs corresponding to these deterministic codes.

We will construct a new deterministic encoder $\check{F} = (\check{f}_1, \check{f}_2, \cdots)$ of the form

$$\check{f}_k(X_1^k) = f^{z_k}_{N_k^{z_k}}(X_{M_1^{z_k}}, X_{M_2^{z_k}}, \cdots, X_{M^{z_k}_{N_k^{z_k}}})$$

The basic idea of our construction is to deterministically and causally split the source sequence X_1^∞ into l subsequences. Since the process is i.i.d., so are these subsequences. Then, we apply the l original deterministic sequential codes to encode these subsequences in a causal way. Finally, we causally reassemble these encoded subsequences into an encoding of the original source sequence. The notation might seem confusing, but this is just the z_k-th original deterministic code operating on the appropriate subsequence of source symbols so far. That appropriate subsequence has $N_k^{z_k}$ elements, which are indexed by $M_1^{z_k}, M_2^{z_k}, \cdots, M^{z_k}_{N_k^{z_k}}$ in the original source.

The key to getting this working is in the "splitting." We must do it in a way that is compatible with P. Without loss of generality, assume that P is such that $\forall 1 \leq i \leq l, p_i > 0$ and $p_i = 0$ otherwise.

Definition 5.1: We call a sequence z_1^∞ a *P-splitting sequence* if $\forall k, 1 \leq z_k \leq l$ and

$$\forall i \lim_{k\to\infty} \frac{\sum_{j=1}^k \delta(z_j - i)}{k} = p_i$$

where δ is the Kronecker delta function.

We let $N_k^i = \sum_{j=1}^k \delta(z_j - i)$. So, N_k^i is the number of times z_j is equal to i up to and including z_k.

But before we proceed, we need to establish that desired z_1^∞ sequences exist.

Lemma 5.1: P-splitting z_1^∞ sequences exist.

There do exist simple algorithms to generate such sequences without needing any randomness at all. In fact, the following simple rule suffices.

(1) Start with the empty sequence. Let $n = 0$
(2) Increment n
(3) Compute N_n^i for all i.
(4) Let j be one for which $(p_j n - N_n^j)$ is maximal.
(5) Set $z_n = j$
(6) Goto 2

It should be clear that the above loop extends sequences in such a way that $\frac{N_n^i}{n}$ converges asymptotically to p_i as n gets large, thereby generating a valid sequence z_1^∞ no matter how ties are broken.

We now use z_1^∞ to construct the set of indices for our subsequences. View N_k^i as a function from non-negative integers k into the non-negative integers. It should be clear that $N_\cdot^i$ is surjective and monotonic by construction since it counts up from zero to infinity. Consider the set valued inverse image $(N_\cdot^i)^{-1}$ and define a new family of sequences $M_k^i = \min\{(N_\cdot^i)^{-1}(k)\}$. This picks out the k-th jump upwards in $N_\cdot^i$, or in other words, the index which corresponds to the k-th occurrence of i in the sequence z_1^∞. Clearly:

- $\forall i, k$ we have $z_{M_k^i} = i$ by construction
- Conversely, if $z_k = i$, then $\exists j$ such that $M_j^i = k$.
- If $i \neq j$ then $\forall k, M_k^i \neq M_k^j$
- $\forall i, M_k^i$ is strictly monotonically increasing in k

These properties assure us that the M^i sequences partition the positive integers in such a way as to define the subsequences that we need. All that remains is to check to assure that our new deterministic encoder is sequential.

Recall that our deterministic encoder is defined by

$$\check{f}_k(X_1^k) = f^{z_k}_{N_k^{z_k}}(X_{M_1^{z_k}}, X_{M_2^{z_k}}, \cdots, X_{M^{z_k}_{N_k^{z_k}}})$$

Sequentiality is checked by verifying that the indices $[M_1^{z_k}, \cdots, M^{z_k}_{N_k^{z_k}}]$ are all less than or equal to k. But since the M_k^i sequences are all strictly monotonically increasing in k it suffices to check the last term: $M^{z_k}_{N_k^{z_k}} = \min\{(N^{z_k})^{-1}(N_k^{z_k})\} = k$, since $N_k^{z_k}$ jumps upward at k because of how it is constructed from z_k. So, by the properties given, the deterministic encoder $\check{F}$ is sequential.

Now, we are ready to state the main result of this section.

Lemma 5.2: $\check{F}$ has rate/distortion performance equal to $\tilde{F}$.

In other words, almost surely we have:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n d(X_k, \check{f}_k(X_1^k)) = \sum_{i=1}^l p_i K_i$$

and

$$\lim_{n\to\infty} \frac{1}{n} H(\check{\hat{X}}_1^n) = \sum_{i=1}^l p_i R_i$$

The above construction establishes that for i.i.d. sources finite randomization does not allow us to reach any additional points of (rate,distortion) performance. $\check{F}$ is a deterministic sequential encoder. So, we can therefore use the existing

result of Neuhoff and Gilbert (Theorem 1.3 above) to show that for sequentially encoding i.i.d. sources, it suffices to consider time-sharing of *memoryless* source encoders.

A. Example of the Performance of Sequential Codes

To see that the Neuhoff and Gilbert theorem implies a fundamental performance gap between the causal and non-causal case, consider the following simple example. Let $\{X_k\}$ be i.i.d. fair coin tosses either 0 or 1. Choose the Hamming distortion measure, namely $\rho(0,0) = \rho(1,1) = 0$ and $\rho(0,1) = \rho(1,0) = 1$. Traditional rate-distortion theory tells us that the non-causal performance achievable is given by $R \geq 1 + K \log K + (1-K)\log(1-K)$ where the r.h.s. equals $R(K)$ as defined in Theorem 1.2 above. However, for this simple source there are exactly four deterministic memoryless source encoders:

(1) $f^1(X) = X$ (Perfect Reconstruction)
(2) $f^2(X) = 0$ (All zeros)
(3) $f^3(X) = 1$ (All ones)
(4) $f^4(X) = \bar{X}$ (Worst Case)

By inspection, f^2 and f^3 have average distortion $K_2 = K_3 = \frac{1}{2}$ and output entropy rate $R_2 = R_3 = 0$. Meanwhile, the perfect reconstruction f^1 has average distortion $K_1 = 0$ while the output entropy rate equals the input entropy rate $R_1 = 1$. The worst case f^4 has average distortion $K_4 = 1$ while also having output entropy rate $R_4 = 1$. So, it is clear that the causally achievable performance region is given by $R_{\text{causal}}(K) \geq 1 - 2K$.

For this binary symmetric source, consider the point of $K = \frac{1}{4}$. Causally, we need at least $R \geq 0.5$. Meanwhile, if we are allowed to look into the future, we only need $R \geq 0.1887$. That is a difference of more than a 264%!

Since $R(K)$ in this example also equals $J^*(K)$, it is clear that the inequality in Theorem 4.1 cannot in general be replaced with an equality. Contrast this with the situation in traditional rate-distortion theory. There, we get equality (look at Theorem 1.2 and let ϵ tend to zero) instead of a lower bound ($R_\infty \geq J_\infty$ in Theorem 4.1). The fact that block coding is allowed (which is non-causal) instead of just sequential coding allows us to generate arbitrary partitions of the space Σ^n as opposed to the "rectangular" partitions forced by sequential coding. This counterexample shows that this geometrical difference in the two schemes leads to this fundamental gap.

VI. Observations

We conclude with some general observations.

1. The finiteness assumption on S can be dropped in so far as our claims that involve only entropy or entropy rate are concerned. Those involving mutual information require the continuity of $I_n(\cdot)$ in two places, Lemma 3.1 and Lemma 3.7. In its absence, the claims concerning $I_n(\cdot)$ in Lemma 3.1 and Lemma 3.7 have to be dropped. The rest of the paper is not affected. If S is countable, $I_n(\cdot)$ can be shown to be lower semicontinuous (This follows from Lemma 5.5.1, p. 122, [9].) and thus Lemma 3.1 still holds in its totality.
2. Let γ denote the random variable as in Lemma 2.5, representing the randomization over deterministic codes. Then
$$\begin{aligned} H(Z^n/X^n) &= H(Z^n/X^n) - H(Z^n/X^n,\gamma) \\ &= I(Z^n;\gamma/X^n) \\ &= H(\gamma/X^n) - H(\gamma/X^n,Z^n) \\ &= H(\gamma) - H(\gamma/X^n,Z^n) \\ &= I(\gamma;X^n,Z^n) \end{aligned} \quad (2)$$
where in the first step we use the fact that Z^n is a function of X^n and γ, and in the last but one step, the independence of γ and X^n. The term $I(\gamma;X^n,Z^n)$ can be interpreted as the redundancy in coding both X^n, Z^n when γ is not known and may be viewed as a measure of 'complexity' of the randomization. Also, since
$$H(Z^n) = I(X^n;Z^n) + I(\gamma;X^n,Z^n),$$
the minimization of the l.h.s. involves a trade-off between the minimization of the two terms on the r.h.s., i.e., between minimizing mutual information and minimizing the 'redundancy of randomization'.
3. Our formulation did not explicitly consider the presence of a noisy channel. As Shannon notes in his paper "Coding Theorems for a Discrete Source with a Fidelity Criterion" [17], the solution to the traditional rate distortion problem corresponds to finding a channel that is just right for the source and allowed distortion level. That is one way of interpreting the fact that the minimization is done over transition probabilities between the source and the reconstruction. On the other hand, the noisy channel coding theorem leads to a source which is just right for the channel since in that case, the maximization is done over input letter probabilities.

An important theorem of traditional rate-distortion theory (Theorem 3 in the above mentioned paper [17]) effectively says that these two solutions can be combined in practical systems. This means that as long as transmission over the channel takes place with a rate (calculated based on the acceptable distortion level) which is less than the capacity of the noisy channel, channel decoding can be done with an arbitrarily small probability of error. This, for a large class of sources, then allows us to achieve an end-to-end distortion that is arbitrarily close to K as long as the channel has $C > R(K)$.

It remains an open question how to incorporate noisy channels into a formulation of a sequential rate-distortion theory.

Conclusions

There is a conceptual issue that has not been dealt with in this paper. If we adopt the definition of sequentiality to mean zero delay, it is unclear that the model investigated in this paper precisely captures communication with zero delay. The correct formulation would be to introduce a source decoder and aggregate the effects of cascading the channel encoder, channel and channel decoder as a fixed finite delay. The

criterion to be adopted for reliable communication would then be to require that the probability of decoding error should asymptotically tend to zero. This would be the analog of the Noisy Channel Coding Theorem for Source Coding.

ACKNOWLEDGEMENT

Support was provided by the Army Research Office under the MURI Grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466, the Department of Defense MURI Grant: Complex Adaptive Networks for Cooperative Control Subaward #03-132, and the National Science Foundation Grant CCR-0325774.

REFERENCES

[1] V. S. BORKAR, White noise representations in stochastic realization theory, SIAM J. Control and Opt. 31 (5), 1993, pp.1093-1102.
[2] V. S. BORKAR, Ergodic control of Markov chains with constraints – the general case, SIAM J. Control and Opt. 32(1), 1994, pp.176-186.
[3] V. S. BORKAR, Probability Theory – An Advanced Course, Springer Verlag, New York, 1995.
[4] V. S. BORKAR, S. K. MITTER, S. TATIKONDA, Optimal sequential vector quantization of Markov sources, submitted.
[5] G. CHOQUET, Lectures on Analysis, Vol. II, W.A. Benjamin, Inc. New York, 1969.
[6] T. M. COVER, J. A. THOMAS, Elements of Information Theory, John Wiley, New York, 1991.
[7] L. E. DUBINS, On extreme points of convex sets, J. Math. Analysis and Appl. 5 (1962), pp.237-244.
[8] L. E. DUBINS, D. FREEDMAN, Measurable sets of measures, Pacific J. Math. 14 (1964), pp. 1211-1222.
[9] R. M. GRAY, Entropy and Information Theory, Springer Verlag, New York, 1990.
[10] R. M. GRAY, D. L. NEUHOFF, J. K. OMURA, Process definitions of distortion-rate functions and source coding theorems, IEEE Trans. on Information Theory IT-21(5), (1975), pp. 524-532.
[11] R. M. GRAY, D. L. NEUHOFF, D. S. ORNSTEIN, Nonblock source coding with a fidelity criterion, Annals of Probability 3(3) (1975), pp. 478-491.
[12] R. M. GRAY, D. L. NEUHOFF, P. C. SHIELDS, A generalization of Ornstein's $\overline{d}$-distance with applications to information theory, Annals of Probability 3 (1979), pp.315-328.
[13] T. S. HAN, An information-spectrum approach to source coding theorems with a fidelity criterion, IEEE Trans. on Information Theory IT-43(4), (1997), pp. 1145-1164.
[14] T. S. HAN, S. VERDU, A general formula for channel capacity, IEEE Trans. on Information Theory IT-39(3), (1993), pp. 752-772.
[15] D. L. NEUHOFF, R. K. GILBERT, Causal source codes, IEEE Trans. on Information Theory IT-28(5), (1982), pp. 701-713.
[16] H. S. WITSENHAUSEN, Some aspects of convexity useful in information theory, IEEE Trans. on Information Theory IT-26(3), (1980), pp. 265-271.
[17] C. E. SHANNON, Collected Papers of C. E. Shannon, eds. N. J. Sloane and A. D. Wyner, IEEE Press, 1993