Rate-Constrained Simulation and Source Coding IID Sources

Mark Z. Mao, *Member, IEEE,* Robert M. Gray, *Life Fellow, IEEE*, and Tamas Linder, *Senior Member, IEEE*

***Abstract*—Necessary conditions for asymptotically optimal sliding-block or stationary codes for source coding and rate-constrained simulation of memoryless sources are presented and used to motivate a design technique for trellis-encoded source coding and rate-constrained simulation. The code structure has intuitive similarities to classic random coding arguments as well as to "fake process" methods and alphabet-constrained methods. Experimental evidence shows that the approach provides comparable or superior performance in comparison with previously published methods on common examples, sometimes by significant margins.**

***Index Terms*—Source coding, simulation, rate-distortion, trellis source encoding**

I. Introduction

ONE view of the basic goal of Shannon source coding with a fidelity criterion or lossy data compression is to covert an information source $\{X_n\}$ into bits which can be decoded into a good reproduction of the original source, ideally the best possible reproduction with respect to a fidelity criterion given a constraint on the rate of transmitted bits. Memoryless discrete-time sources have long been a standard benchmark for testing source coding or data compression systems. Although of limited interest as a model for real world signals, independent identically distributed (IID) sources provide useful comparisons among different coding methods and designs. In addition, specific examples such as Gaussian and uniform sources can provide intuitive interpretations of how coding schemes yield good performance and they can serve as building blocks for more complicated processes such as linear models driven by IID processes.

A separate, but intimately related, topic is that of rate-constrained simulation — given a "target" random process such as an IID Gaussian process, what is the *best* possible imitation of the process that can be generated by coding a simple discrete IID process with a finite entropy rate? Here "best" can be quantified by a metric on random processes such as the generalized Ornstein $\overline{d}$ distance (or Monge-Kantorovich transportation distance/ Wasserstein distance extended to random processes). For example, what is the best imitation Gaussian process with only one bit per symbol?

M. Z. Mao and R. M. Gray are with the Department of Electrical Engineering, Stanford University, Stanford, CA, 94305, USA. Tamas Linder is with the Department of Mathematics and Statistics, Queens University, Kingston, Ontario, Canada K7L 3N6, e-mail: markmao at stanford.edu, rmgray at stanford.edu, linder at mast.queensu.ca.

Research partially supported by the NSF under grant CCF-0846199-000 and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

December 14, 2009.

The two problems of source coding and simulation are intimately related, as is obvious from the general block diagram of Fig. 1. The right half of the source coding system

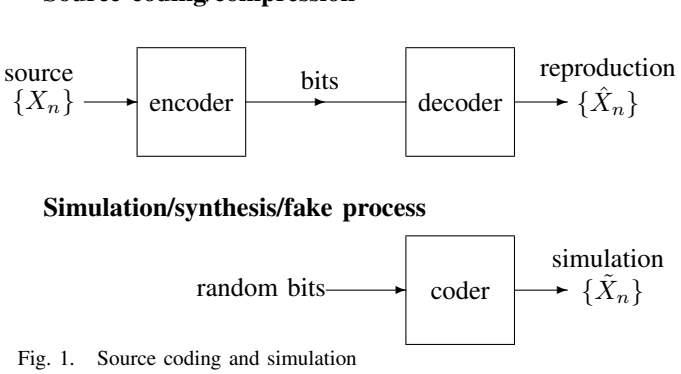

Fig. 1. Source coding and simulation

of Fig. 1 resembles the simulation system and the picture suggests that designing a good decoder for source coding is equivalent to designing a good simulator. Intuitively (and mathematically [10], [12]), if the source code operates near the Shannon optimum, then the bits being transmitted should be close to an IID equiprobable Bernoulli process. Since the reproduction process is close to the input process, it should be a good simulation given the bit constraint if the channel bits from the source are replaced by coin flips. Conversely, if one has a good simulation, in theory one should be able to construct good codewords for long sequences of source inputs by finding the best match between the input sequence and the possible simulator output sequences. Hence it is natural to suspect that the optimal performance for each system with a common rate constraint and fidelity criterion should be the same and that good codes for either system can be constructed from those for the other.

Rigorous results along this line were developed in 1977 [8] showing that the two optimization problems are equivalent and optimal (or nearly optimal) source coders imply optimal (or nearly optimal) simulators and vice versa for the specific case of stationary codes (shifting the input results in a corresponding shift of the output) and sources that are B-processes (stationary codings of IID processes).

Similar results for asymptotically long block codes were developed for more general sources by Steinberg and Verdu in 1996 [34], but our focus is on stationary codes and the behavior of processes rather than on the asymptotics of finite-dimensional distributions, which might not correspond to the joint distributions of a stationary process.

We introduce a design technique for trellis-encoded source coding that combines the goals of the fake process viewpoint with the usual random coding argument for trellis codes and provides better performance than either individually. The proposed scheme also resembles more recent variations on the old methods. Unlike previous work [39], [24], [35], [3], rather than formulate intuitive guidelines for good code design we prove several necessary conditions which optimal or asymptotically optimal source codes must satisfy. These include Pearlman's observation [24] that the marginal reproduction distribution should approximate the Shannon optimal reproduction and that the reproduction process should be approximately white. We give a code construction which provably satisfies a key necessary condition and which is shown experimentally to satisfy the other necessary conditions while providing performance comparable to or superior to previously published work. In particular, the performance of our trellis source encoder has not yet reached a plateau and seems to approach the Shannon limit with increasing encoder complexity.

The rest of the paper is organized as follows. In Section II we give an overview of definitions and concepts we need for stating our results and in Section III we state and prove the necessary conditions for optimum trellis-encoded source code design. Section IV introduces the new design technique and Section V presents experimental results for encoding memoryless Gaussian, uniform, and Laplacian sources.

II. Preliminaries

A. A note on notation

We deal with random objects which will be denoted by capital letters. These include random variables X_n, N-dimensional random vectors $X^N = (X_0, X_1, \ldots, X_{N-1})$, and random processes $\{X_n; n \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of all integers. The generic notation X might stand for any of these random objects, where the specific nature will either be clear from context or stated (this is to avoid notational clutter when possible). Lower case letters will correspond to sample values of random objects. For example, given an alphabet A (such as the real line $\mathbb{R}$ or the binary alphabet $\{0,1\}$), then a random variable X_n may take on values $x_n \in A$, an N-dimensional random vector X^N may take on values $x^N \in A^N$, the Cartesian product space, and a random process $\{X_n; n \in \mathbb{Z}\}$ may take on values $\{x_n; n \in \mathbb{Z}\} = (\cdots, x_{-1}, x_0, x_1, \cdots) \in A^\infty$. A lower case letter without subscript or superscript may stand for a member of any of these spaces, depending on context.

B. Stationary and sliding-block codes

A stationary or sliding-block code can be viewed as a time-invariant filter, in general nonlinear. It operates on an input sequence to produce an output sequence in such a way that shifting the input sequence results in a shifted output sequence. In the abstract, a stationary code $\bar{f}$ with an input alphabet A (typically $\mathbb{R}$ or a Borel subset for an encoder or $\{0,1\}$ for a decoder) and output alphabet B (typically $\{0,1\}$ for an encoder or some subset $\mathbb{R}$ for a decoder) is a measurable mapping (with respect to suitable σ-fields) of an infinite input sequence (in A^∞) into an infinite output sequence (in B^∞) with the property that $\bar{f}(T_A x) = T_B \bar{f}(x)$, where T_A is the (left) shift on A^∞, that is, $T_A(\cdots, x_{-1}, x_0, x_1, \cdots) = (\cdots, x_0, x_1, x_2 \cdots)$. The sequence-to-sequence mapping from $\bar{f} : A^\infty \to B^\infty$ is described by the sequence-to-symbol mapping defined by code output at time 0, $f(x) = \bar{f}(x)_0$ since $\bar{f}(x)_n = \bar{f}(T_A^n x)_0 = f(T_A^n x)$. More concretely, the sequence-to-symbol mapping f usually depends on only a finite window of the data, in which case the output random process, say $\{Y_n\}$, can be expressed as a mapping on the contents of a shift register containing samples of the input random process $\{X_n\}$ as depicted in Fig. 2. Note that the time order is reversed in the shift-register representation since in the shift register new input symbols flow in from the left and exit from the right, but the standard way of writing a sequence is $\cdots, X_{n-2}, X_{n-1}, X_n, X_{n+1}, X_{n+2}, \cdots$ with "past" symbols on the left and "future" symbols on the right. The *length* of the code is the length of the shift register or dimension of the vector argument, $L = N_1 + N_2 + 1$. Both f and $\bar{f}$ will

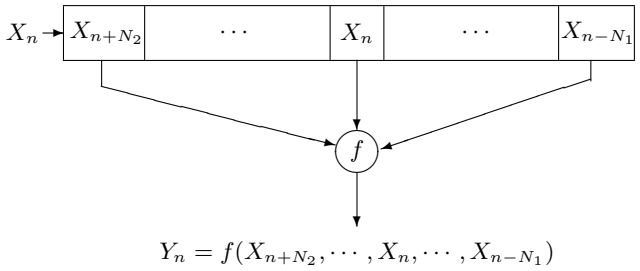

Fig. 2. Stationary or sliding-block code

be referred to as stationary or sliding-block codes. Stationary codes (unlike traditional block codes) are well-defined for infinite length codes, that is, codes are well-defined for which an infinite input sequence is viewed to produce a single output symbol.

Unlike block codes, stationary codes preserve statistical characteristics of the coded process, including stationarity, ergodicity, and mixing. If a stationary and ergodic source $\{X_n\}$ is encoded into bits by a stationary code f, which are in turn decoded into a reproduction process $\{\hat{X}_n\}$ by another stationary code g, then the resulting pair process $\{X_n, \hat{X}_n\}$ and output process $\{\hat{X}_n\}$ are also stationary and ergodic.

Given any block code, a stationary code with similar properties can be constructed (at least in theory) and vice versa. Thus good codes of one type can be used to construct good codes of the other (at least in theory) and the optimal performance for the two classes of codes is the same [29], [14], [7], [9].

C. Fidelity and distortion

As a fidelity criterion for source coding system we will use a *single-letter* or additive distortion measure defined in terms of a nonnegative distortion measure $d(x, y)$, $x \in A, y \in \hat{A}$. The distortion incurred when a block of source letters $x^N \in A^N$ is reproduced as $y^N \in \hat{A}^N$ is

$$d_N(x^N, y^N) = \sum_{i=0}^{N-1} d(x_i, y_i).$$

Given random vectors X^N, Y^N with a joint distribution π^N, the average distortion is defined by the expectation $d_N(\pi^N) = E[d_N(X^N, Y^N)]$.

Given a stationary pair process $\{X_n, Y_n\}$, the average distortion between N-tuples is given by the single-letter characterization $N^{-1}E[d_N(X^N, Y^N)] = E[d(X_0, Y_0)] = d_1(\pi^1)$ and hence a measure of the fidelity (or, rather, lack of fidelity) of a stationary coding and decoding of a stationary source X_n into a reproduction $\hat{X}_n$ is the average distortion

$$D(f, g) = E[d(X_0, \hat{X}_0)].$$

The emphasis in this paper will be the case where $\hat{A} \subset A = \mathbb{R}$ and the distortion is the common squared error distortion, $d(x, y) = (x - y)^2$ (so that the distortion on n-tuples is the square of the ℓ_2 distance). Also of interest is the Hamming distortion, where $d(x, y) = 0$ if $x = y$ and 1 otherwise.

D. Optimal source coding

Let $\mathcal{C}(A, B)$ denote the collection of all sliding-block codes with input alphabet A and output alphabet B. The operational distortion-rate function is defined by

$$\delta_X(R) = \inf_{f \in \mathcal{C}(A,B), g \in \mathcal{C}(B,\hat{A})} D(f, g).$$

E. Distance measures for random vectors and processes

A distortion measure d induces a natural notion of a "distance" between random vectors and processes (the quotes will be removed when the relation to a true distance or metric is clarified). The basic idea has many names and a wide literature spanning many fields. The *optimal transportation cost* between two probability distributions, say μ_X and μ_Y corresponding to random variables (or vectors) defined on a common (Borel) probability space $(A, \mathcal{B}(A))$ with a nonnegative cost function d is defined as

$$\mathcal{T}(\mu_X, \mu_Y) = \inf_{\pi \in \mathcal{P}(\mu_X, \mu_Y)} E_\pi d(X, Y),$$

where $\mathcal{P}(\mu_X, \mu_Y)$ is the class of all probability distributions on $(A, \mathcal{B}(A))^2$ having μ_X and μ_Y as marginals, that is, $\pi(F \times A) = \mu_X(F)$, $\pi(A \times F) = \mu_Y(F)$ for all $F \in \mathcal{B}(A)$. Any π having the prescribed marginals μ_X and μ_Y is called a *coupling* of μ_X and μ_Y. The reader is referred to Villani [37] and Rachev and Rüschendorf [26] for extensive development and references. The most important special case is when the cost function is a nonnegative power of an underlying metric

$$d(x, y) = m(x, y)^r \tag{1}$$

where A is a complete, separable metric (Polish) space with respect to m. In this case $\mathcal{T}(\mu_X, \mu_Y)^{\min(1,1/r)}$ is a metric. The case $r = 0$ is used to denote the Hamming distance. The notation $\mathcal{T}_2$ and $\mathcal{T}_0$ will be used to denote the two most important cases of the optimal transportation cost with respect to the squared error and Hamming distance, respectively.

Given two stationary processes with process distributions μ_X and μ_Y, let μ_{X^N} and μ_{Y^N} denote the induced N-dimensional distributions for all positive integers N. Let d_N be a single-letter fidelity criterion induced by $d(x, y)$, $x, y \in A$. Define the (generalized) $\overline{d}$ distance [13]

$$\overline{d}(\mu_X, \mu_Y) = \sup_N N^{-1}\mathcal{T}(\mu_{X^N}, \mu_{Y^N}).$$

If d is a metric, then so is $\overline{d}$. If d is the Hamming metric, this is Ornstein's d-bar distance [22], [23]. If d is a power of an underlying metric as in (1), then $\overline{d}(\mu_X, \mu_Y)^{\min(1,1/r)}$ will also be a metric. We will refer to $\overline{d}$ as the "$\overline{d}$-distance" whether or not it is actually a true metric. We distinguish the most important cases by subscripts, in particular $\overline{d}_2$ denotes $\overline{d}$ with d squared error (and hence $\sqrt{\overline{d}_2}$ is a metric) and $\overline{d}_0$ denotes $\overline{d}$ with d equal to the Hamming distance ($\overline{d}_0$ is a metric).

The basic definition for the $\overline{d}$ distances is in terms of a supremum over vector optimizations, but for stationary processes there is a simpler characterization:

$$\overline{d}(\mu_X, \mu_Y) = \inf_{\pi \in \mathcal{P}(\mu_X, \mu_Y)} E_\pi[d(X_0, Y_0)] \tag{2}$$

where the infimum is over all stationary processes (or stationary and ergodic processes if μ_X and μ_Y are ergodic). This and many other properties of the $\overline{d}$ and generalized $\overline{d}$ are detailed in [22], [23], [13], [11]. Properties relevant here include the following:

1) For stationary processes,
$$\overline{d}(\mu_X, \mu_Y) = \lim_{N \to \infty} N^{-1}\mathcal{T}(\mu_{X^N}, \mu_{Y^N}). \tag{3}$$
2) If the processes are both IID, then
$$\overline{d}(\mu_X, \mu_Y) = \mathcal{T}(\mu_{X_0}, \mu_{Y_0}), \tag{4}$$
and hence the $\overline{d}$-distance reduces to the transportation distance between the one dimensional marginal distributions.
3) If the processes are both stationary and ergodic, the distance can be expressed as the infimum over the limiting distortion between any two frequency-typical sequences of the two processes. Thus the $\overline{d}$-distance between the two processes is the amount by which a frequency-typical sequence of one process must be changed in a time average d sense to produce a frequency-typical sequence of another process.

The $\overline{d}$ process distance can be used to characterize both the optimal source coding and the optimal rate-constrained simulation problem. Let $\{X_n\}$ be a random process described by a process distribution μ_X and let $\{Z_n\}$ be an IID random process with alphabet B of size $\|B\| = 2^R$ and distribution μ_Z. The optimal simulation of the process $X = \{X_n\}$ with process distribution μ_X given the process $Z = \{Z_n\}$ with process distribution μ_Z is characterized by

$$\Delta_{X|Z}(R) = \inf_{f \in \mathcal{C}(B,\hat{A})} \overline{d}(\mu_X, \mu_{\bar{f}(Z)}) \tag{5}$$

where $\mu_{\bar{f}(Z)} = \mu_Z \bar{f}^{-1}$ is the process distribution resulting from a stationary coding of Z using f, i.e., for all events F $\mu_{\bar{f}(Z)}(F) = \mu_Z(\bar{f}^{-1}(F))$.

F. Entropy rate

Alternative characterizations of the optimal source coding and simulation performance can be stated in terms of the entropy rate of a random process. As we will be dealing with both discrete and continuous alphabet processes and with some borderline processes that have continuous alphabets yet finite entropy, suitably general notions of entropy as found in mathematical information theory and ergodic theory are needed (see, e.g.., [25], [22], [23], [11]). For a finite-alphabet random process, define as usual the Shannon entropy of a random vector or, equivalently, of its distribution by $H(X^N) = H(\mu_{X^N}) = -\sum_{x^N} \mu_{X^N}(x^N) \log \mu_{X^N}(x^N)$ and the Shannon entropy rate of the process X by $H(X) = H(\mu_X) = \inf_N N^{-1} H(X^N)$. If the process is stationary, then

$$H(X) = \lim_{N\to\infty} N^{-1} H(X^N). \tag{6}$$

In the general case of a continuous alphabet, the entropy rate is given by the Kolmogorov-Sinai invariant $H(X) = \sup_f H(\mu_{\overline{f}(X)})$, where the supremum is over all finite-alphabet stationary codes. In other words, the entropy rate of a continuous alphabet random process is the supremum of the ordinary Shannon entropy rates of finite-alphabet coded versions of the process. It is important to note that (6) need not hold when the alphabet is not finite and that a random process with a continuous alphabet can have an infinite finite-order entropy and a finite entropy rate.

G. Constrained entropy rate optimization

A stationary and ergodic process is called a B-process if it is obtained by a stationary coding of an IID process. If the source is stationary and ergodic, then [8]

$$\Delta_{X|Z}(R) = \inf_{B\text{-processes } \nu: H(\nu)\le R} \overline{d}(\mu_X, \nu), \tag{7}$$

that is, the best simulation by coding coin flips in a stationary manner is the best simulation of X by any B-process having entropy rate R bit per symbol or less. If X were itself discrete and a B-process with entropy rate less than or equal to R, then Ornstein's isomorphism theorem [22], [23] (or the weaker Sinai-Ornstein theorem) implies that $\Delta_{X|Z}(R) = 0$. In words, a B-process can be stationarily encoded into any other B process having equal or smaller entropy rate.

The $\overline{d}$-distance also yields a characterization of the operational distortion rate function [14]:

$$\delta_X(R) = \inf_{\nu: H(\nu)\le R} \overline{d}(\mu_X, \nu), \tag{8}$$

where the infimum is over all stationary and ergodic processes. Comparing (7) and (8), obviously $\Delta_{X|Z}(R) \ge \delta_X(R)$. If the source X is also a B-process, then the two infima are the same and $\Delta_{X|Z}(R) = \delta_X(R)$.

A related operational distortion-rate function resembling the simulation problem replaces the encoder/decoder with a common alphabet by a single code into a reproduction having a constrained entropy rate. Suppose that a source X is encoded by a sliding-block code f directly into a reproduction $\hat{X}$ with process distribution $\mu_{\hat{X}} = \mu_{\overline{f}(X)}$. What coding yields the smallest distortion under the constraint that the output entropy rate is less than or equal to R? In this case, unsurprisingly

$$\inf_{f\in\mathcal{C}(A,\hat{A}): H(\mu_{\overline{f}(X)})\le R} E[d(X_0, \hat{X}_0)] = \delta_X(R). \tag{9}$$

These relations implicitly define optimal codes and optimal performance, but they do not say how to evaluate the optimal performance or design the codes for a particular source. The Shannon rate-distortion function solves the first problem.

H. Shannon rate-distortion functions

In the discrete alphabet case the Nth order average mutual information between random vectors X^N and Y^N is given by $I(X^N, Y^N) = H(X^N) + H(Y^N) - H(X^N, Y^N)$. In general $I(X^N, Y^N)$ is given as the supremum of the discrete alphabet average mutual information over all possible discretizations or quantizations of X^N and Y^N. If the joint distribution of X^N and Y^N is π^N, then we also write $I(\pi^N)$ for $I(X^N, Y^N)$.

The Shannon rate-distortion function [28] is defined by

$$R_X(D) = \inf_N N^{-1} R_{X^N}(D) = \lim_{N\to\infty} N^{-1} R_{X^N}(D)$$

$$R_{X^N}(D) = \inf_{\pi^N: \pi^N\in\mathcal{P}(\mu_{X^N}), N^{-1}Ed_N(\pi^N)\le D} N^{-1} I(\pi^N) \tag{10}$$

where the infimum is over all joint distributions π^N for X^N, Y^N with first marginal distribution μ_{X^N} ($\pi^N \in \mathcal{P}(\mu_{X^N})$) and $N^{-1}E[d(X^N, Y^N)] \le D$. The dual function is the Shannon distortion rate function:

$$D_X(R) = \inf_N N^{-1} D_{X^N}(R) = \lim_{N\to\infty} N^{-1} D_{X^N}(R)$$

$$D_N(R) = \inf_{\pi^N: \pi^N\in\mathcal{P}(\mu_{X^N}), N^{-1}I(\pi^N)\le R} N^{-1} Ed_N(X^N, Y^N).$$

Source coding theorems show that under suitable conditions $\delta_X(R) = D_X(R)$. (See, e.g., [14], [7], [9] for source coding theorems for stationary codes.)

I. Evaluation of rate-distortion functions

Csiszár [2] provided quite general versions of Gallager's [5] Kuhn-Tucker optimization for evaluating the rate-distortion function. We here state the relevant results in a form and generality useful for our purposes. We consider finite-order rate-distortion functions for vectors $X = X^N$. The following lemma is a combination of Lemma 1.2, corollary to Lemma 1.3, and equations (1.11) and (1.15) in Csiszár [2].

Lemma 1: If $R_X(D) < \infty$, then

$$R_X(D) = \max_{\beta\ge 0} (\Gamma(\beta) - \beta D) \tag{11}$$

$$\Gamma(\beta) = \inf_{\pi\in\mathcal{P}(\mu_X)} (I(\pi) + \beta d(\pi)) \tag{12}$$

$$= \inf_{\mu_Y} \int d\mu_X(x) \log \frac{1}{\int d\mu_Y(y) e^{-\beta d(x,y)}} \tag{13}$$

where the final line defines $\Gamma(\beta)$ as an infimum over all distributions on $\hat{A}$. There exists a value β such that the straight

line of slope $-\beta$ is tangent to the rate-distortion curve at $(R(D), D)$, in which case β is said to be *associated with* D. If π achieves a minimum in (12), then $D = d(\pi), R(D) = I(\pi)$.

Thus for a given D there is a value of β associated with D, and for this value the evaluation of the rate-distortion curve can be accomplished by an optimization over all distributions μ_Y on the reproduction alphabet. If a minimizing π exists, then the resulting marginal distribution for μ_Y is called a *Shannon optimal reproduction distribution*. In general this distribution need not be unique.

The next lemma shows that under the assumptions of a distortion measure that is a power of a metric derived from a norm, there exists a π achieving the minimum of (10) and hence also a Shannon optimal reproduction distribution. Both the lemma and the subsequent corollary are implied by the proof of Csiszár's Theorem 2.2 and the extension of the reproduction space from compact metric to Euclidean spaces discussed at the bottom of p. 66 of [2]. In the corollary, the roles of distortion and mutual information are interchanged to obtain the distortion-rate version of the result.

Lemma 2: Given a random vector X with an alphabet A which is a finite-dimensional Euclidean space with norm $\|x\|$, a reproduction alphabet $\hat{A} = A$, and a distortion measure $d(x, y) = \|x-y\|^r$, $r > 0$, then there exists a distribution π on $A \times A$ achieving the the minimum of (16). Hence a Shannon N-dimensional optimal reproduction distribution exists for the Nth order rate-distortion function.

Corollary 1: Given the assumptions of the lemma, suppose that $\pi^{(n)}$, $n = 1, 2, \ldots$ is sequence of distributions on $A \times \hat{A}$ with marginals μ_X and $\mu_{Y^{(n)}}$ for which for $n = 1, 2, \ldots$

$$I(\pi^{(n)}) = I(X, Y^{(n)}) \le R, \tag{14}$$

$$\lim_{n\to\infty} E[d(X, Y^{(n)})] = D_X(R). \tag{15}$$

Then $\mu_{Y^{(n)}}$ has a subsequence that converges weakly to a Shannon optimal reproduction distribution. If the Shannon distribution is unique, then $\mu_{Y^{(n)}}$ converges weakly to it.

J. IID sources

If the process X is IID, then

$$R_X(D) = R_{X_0}(D) = \inf_{\pi:\pi\in\mathcal{P}(\mu_{X_0}), Ed(X_0,Y_0)\le D} I(X_0, Y_0). \tag{16}$$

If a Shannon optimal distribution exists for the first-order rate distortion-function, then this guarantees that it exists for all finite-order rate-distortion functions and that the optimal N-th order distribution is simply the product distribution of N copies of the first-order optimal distribution.

Rose [27] makes two observations important to this paper in his development of a mapping approach to rate-distortion computation and analysis. First, he argues that the minimization of (13) over all distributions on the reproduction alphabet can be replaced by a minimization over all measurable mappings from the unit interval into the reproduction alphabet:

$$\Gamma(\beta) = \inf_f \int d\mu_{X_0}(x) \log \frac{1}{\int_0^1 d\lambda(u) e^{-\beta d(x, f(u))}} \tag{17}$$

where λ denotes the Lebesgue measure on $(0, 1)$ (the uniform distribution on the unit interval). He then provides an alternative to the traditional implementation of the optimization of Lemma 1 for $N = 1$ using the Blahut algorithm [1] by an optimization over mappings using a form of annealing. A significant advantage of his approach is that it allows one to avoid the discretization of the input alphabet of a continuous random variable (which can affect the theory and evaluation of the rate distortion function) and instead shifts the focus to the reproduction alphabet. In particular, Rose proves that for a continuous input random variable and the squared error distortion, the Shannon optimal reproduction distribution will be (absolutely) continuous only in the special case where the Shannon lower bound to the rate distortion function holds with equality, e.g., in the case of a Gaussian source and squared error distortion. In other cases, the optimum reproduction distribution is discrete, and for source distributions for bounded support (e.g., the uniform $(0, 1)$ source), the Shannon optimal reproduction distribution will have finite support, that is, it will be describable by a probability mass function (PMF) with with a finite domain. Rose's algorithm attempts to find this finite alphabet for a given value of the parameter β. Application of the Rose algorithm will convert a possibly continuous source into a specific finite alphabet optimal reproduction alphabet directly, avoiding the indirect path of first discretizing the input distribution and then performing a discrete Blahut algorithm, which is the approach inherent to the constrained alphabet rate-distortion theory and code design algorithm of Finamore and Pearlman [4]. There is no proof that Rose's annealing algorithm actually converges to the optimal solution, but our numerical results support his arguments.

K. Trellis encoding

A sliding-block code can be represented as a shift register into which input symbols are shifted together with a function mapping the shift register contents into a single symbol in the output alphabet as depicted in general in Fig. 2. If the decoder of a source coding system is a finite-length sliding-block code, then encoding can be accomplished using a Viterbi algorithm search of the trellis diagram "colored" (or "labeled" or "populated") by the decoder. The trellis is a directed graph showing the action of a finite-state machine with all but the leftmost (newest) symbol in the shift register constituting the state and the leftmost symbol in the figure being the input. Branches connecting each state are labeled by the output (or an index for the output in a reproduction codebook) produced by receiving a specific input (upper branch) in a given state.

In a source coding system the trellis can be searched by a Viterbi algorithm to find the best path map. The Viterbi algorithm yields a block encoder matched to the sliding-block decoder, but it can be used to construct a stationary encoder by using standard techniques from ergodic theory to convert a block code into a sliding-block code with similar properties. In practice, however, the Viterbi algorithm can simply be used as a block encoder with reasonable complexity. A source coding system having this form is a *trellis source encoding system*. As described, the trellis and the corresponding trellis

branch labels are time-invariant. In particular, the sliding-block decoder does not change with time and every level of the trellis has identical branch labels. The original 1974 source coding theorem for trellis encoded IID sources [38] was proved for time-varying codes by using a variation of Shannon random coding, successive levels of the trellis were labeled randomly based on IID random variables chosen according to the test channel output distribution arising in the evaluation of the Shannon rate-distortion function.

Early research on trellis encoding design was mostly concerned with time-varying trellises with randomly generated labels, reflecting the structure of the coding theorem. In particular, Wilson and Lytle [39] populated their trellis using IID random labels chosen according to the Shannon optimal reproduction distribution. A later source coding theorem for time-invariant trellis encoding [8] was based on the sliding-block source coding theorem [14], [7] and was purely an existence proof; it did not suggest any implementable design techniques. Two early techniques for time-invariant code design were the fake process design [16] and a Lloyd clustering approach conditioned on the shift register states [32], [33]. The former technique was based on a heuristic argument involving optimal simulation and the $\overline{d}$-distance formulation of the operational distortion rate function. The idea was to color a trellis with a process as close in $\overline{d}$ as possible to the original source. While the goal is correct, the heuristic adopted to accomplish it was flawed: the design attempted to match the marginal distribution and the power spectral density of the reproduction with those of the original source. As pointed out by Pearlman [24] and proved in this paper, the marginal distribution of the trellis labels should instead match the Shannon optimal distribution, not the original source distribution.

Pearlman's theoretical development [24] was based on his and Finamore's constrained-output alphabet rate-distortion [4], which involved a prequantization step prior to to designing a trellis encoder for the resulting finite-alphabet process. Pearlman provided a coding theorem and an implementation for a time-invariant trellis encoding, but used the artifice of a subtractive dithering sequence to ensure the necessary independence of successive trellis branch labels over the code ensemble. Because of the dithering, the overall code is not time-invariant.

Marcellin and Fisher in 1990 [17] introduced trellis-coded quantization (TCQ) based on an analogy with coded modulation in the dual problem of trellis decoding for noisy channels. The intuition was spot on for coding uniform IID processes, and provided a coding technique of much reduced complexity that has since become one of the most popular compression systems for a variety of signals. The dual code argument is strong only for the uniform case, but variations of the idea have proved quite effective in a variety of systems. TCQ has a default assignment of reproduction values to trellis branches using a Lloyd-optimized quantizer, but the levels can also be optimized.

Some techniques, including TCQ in our experiments, tend to reach a performance "plateau" in that performance improvement with complexity becomes negligible well before the complexity becomes burdensome. In TCQ this can be attributed to constraints placed on the system to ensure low complexity. The technique introduced here has not (yet) shown any such plateau.

More recently, van der Vleuten and Weber [35] combined the fake process intuition with TCQ to obtain improved trellis coding systems for IID sources. They incorrectly stated that [16] had shown that a necessary condition for optimality for trellis reproduction labels for coding an IID source is that the reproduction process be uncorrelated (white) when the branch labels are chosen in an equiprobable independent fashion. This is indeed an intuitively desirable property and it was used as a guideline in [16] — but it was not shown to be necessary. Eriksson et al. [3] used linear congruential (LC) recursions to generate trellis labels and reproduction values to develop the best codes of the time for IID sources to date by establishing a set of "axioms" of desirable properties for good codes (including a flat reproduction spectrum) and then showing that a trellis decoder based on an inverse CDF of a sequence produced by linear recursion relations meets the conditions. Because of the CDF matching and spectral control, the system can also be viewed as a variation on the fake process approach. Eriksson et al. observe that a problem with TCQ is the constrained ability to increase alphabet size for a fixed rate and they argue that larger alphabet size can always help. This is not correct in general, although it is for the Gaussian source where the Shannon optimal distribution is continuous. For other sources, such as the uniform, the Shannon optimal has finite support and optimizing for an alphabet that is too large or not the correct one will hurt in general. As with TCQ, the approach allowed optimization of the reproduction values assigned to trellis branch labels.

III. Necessary Conditions for Optimal and Asymptotically Optimal Codes

A sliding-block code (f, g) for source coding is said to be optimum if it yields an average distortion equal to the operational distortion-rate function, $D(f, g) = \delta_X(R)$. Unlike the simple scalar quantizer case (or the nonstationary vector quantizer case), however, there are no simple conditions for guaranteeing the existence of an optimal code. Hence usually it is of greater interest to consider codes that are asymptotically optimal in the sense that their performance approaches the optimal in the limit, but there might not be a code which actually achieves the limit. More precisely, a sequence of rate-R sliding-block codes f_n, g_n, $n = 1, 2, \ldots$, for source coding is *asymptotically optimal* (a.o.) if

$$\lim_{n\to\infty} D(f_n, g_n) = \delta_X(R) = D_X(R). \tag{18}$$

An optimal code (when it exists) is trivially asymptotically optimal and hence any necessary condition for an asymptotically optimal sequence of codes also applies to a fixed code that is optimal by simply equating every code in the sequence to the fixed code.

Similarly, a simulation code g is optimal if $\overline{d}(\mu_X, \mu_{\bar{g}(Z)}) = \Delta(X|Z)$ and a sequence of codes g_n is asymptotically optimal if

$$\lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\bar{g}_n(Z)}) = \Delta(X|Z). \tag{19}$$

A. Process approximation

The following lemma provides necessary conditions for asymptotically optimal codes which are a slight generalization and elaboration of Theorem 1 of Gray and Linder [12]. A proof is provided in the Appendix.

Lemma 3: Given a real-valued stationary ergodic process X, suppose that f_n, g_n $n = 1, 2, \ldots$ is an asymptotically optimal sequence of stationary source codes for X with encoder output/decoder input alphabet B of size $\|B\| = 2^R$ for integer rate R. Denote the resulting reproduction processes by $\hat{X}^{(n)}$ and the B-ary encoder output/decoder input processes by $U^{(n)}$. Then

$$\begin{aligned} \lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\hat{X}^{(n)}}) &= D_X(R) \\ \lim_{n\to\infty} H(\hat{X}^{(n)}) &= \lim_{n\to\infty} H(U^{(n)}) = R \\ \lim_{n\to\infty} \bar{d}_0(U^{(n)}, Z) &= 0, \end{aligned}$$

where Z is an IID equiprobable process with alphabet size 2^R.

These properties are quite intuitive:

- The process distance between a source and an approximately optimal reproduction of entropy rate less than R is close to the Shannon distortion rate function. Thus frequency-typical sequences of the reproduction should be as close as possible to frequency-typical source sequences.
- The entropy rate of an approximately optimal reproduction and of the resulting encoded B-ary process must be near the maximum possible value.
- The sequence of encoder output processes approaches an IID equiprobable source in the Ornstein process distance. If $R = 1$, the encoder output bits should look like fair coin flips.

If X is a B-process, then a sequence of a.o. simulation codes g_n yielding a reproduction processes $\tilde{X}^{(n)}$ satisfies $\lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\tilde{X}^{(n)}}) = \Delta_{X|Z}(R) = D_X(R)$ and a similar argument to the proof of the previous lemma implies that $\lim_{n\to\infty} H(\hat{X}^{(n)}) = H(Z) = R$.

B. Moment conditions

The next set of necessary conditions concerns the squared error distortion and resembles a standard result for scalar and vector quantizers (see, e.g., [6], Lemmas 6.2.2 and 11.2.2). The proof differs, however, in that in the quantization case the centroid property is used, while here simple ideas from linear prediction theory accomplish a similar goal. Define in the usual way the covariance $\mathrm{COV}(X, Y) = E[(X - E(X))(Y - E(Y))]$.

Lemma 4: Given a real-valued stationary ergodic process X, suppose that If f_n, g_n is an asymptotically optimal sequence of codes (with respect to squared error) yielding reproduction processes $\hat{X}^{(n)}$ with entropy rate $H(\hat{X}) \le R$, then

$$\lim_{n\to\infty} E(\hat{X}_0^{(n)}) = E(X_0) \tag{20}$$

$$\lim_{n\to\infty} \frac{\mathrm{COV}(X_0, \hat{X}_0^{(n)})}{\sigma^2_{\hat{X}_0^{(n)}}} = 1 \tag{21}$$

$$\lim_{n\to\infty} \sigma^2_{\hat{X}_0^{(n)}} = \sigma^2_{X_0} - D_X(R) \tag{22}$$

Defining the error as $\epsilon_0^{(n)} = \hat{X}_0^{(n)} - X_0$, then the necessary conditions become

$$\lim_{n\to\infty} E(\epsilon_0^{(n)}) = 0 \tag{23}$$

$$\lim_{n\to\infty} E(\epsilon_0^{(n)} \hat{X}_0^{(n)})) = 0 \tag{24}$$

$$\lim_{n\to\infty} \sigma^2_{\epsilon_0^{(n)}} = D_X(R). \tag{25}$$

The results are stated for time $k = 0$, but stationarity ensures that they hold for all times k.

Proof: For any encoder/decoder pair (f_n, g_n) yielding a reproduction process $\hat{X}^{(n)}$

$$\begin{aligned} D(f_n, g_n) &\ge \inf_{a,b\in\mathbb{R}} D(f_n, ag_n + b) \\ &\ge D_X(R) = \inf_{f,g} D(f, g) \end{aligned}$$

where the second inequality follows since scaling a sliding-block decoder by a real constant and adding a real constant results in another sliding-block decoder with entropy rate no greater than that of the input. The minimization over a and b for each n is solved by standard linear prediction techniques as

$$a_n = \frac{\mathrm{COV}(X_0, \hat{X}_0^{(n)})}{\sigma^2_{\hat{X}_0^{(n)}}} \tag{26}$$

$$b_n = E(X_0) - a_n E(\hat{X}_0^{(n)}), \tag{27}$$

$$\begin{aligned} \inf_{a,b} D(f_n, ag_n + b) &= D(f_n, a_n g_n + b_n) \\ &= \sigma^2_{X_0} - a_n^2 \sigma^2_{\hat{X}_0^{(n)}}. \end{aligned} \tag{28}$$

Combining the above facts we have that since (f_n, g_n) is an asymptotically optimal sequence,

$$\begin{aligned} D_X(R) &= \lim_{n\to\infty} D(f_n, g_n) \ge \lim_{n\to\infty} D(f_n, a_n g_n + b_n) \\ &\ge D_X(R) \end{aligned} \tag{29}$$

and hence that both inequalities are actually equalities. The final inequality (29) being an equality yields

$$\lim_{n\to\infty} a_n^2 \sigma^2_{\hat{X}_0^{(n)}} = \sigma^2_{X_0} - D_X(R). \tag{30}$$

Application of asymptotic optimality and (26) to

$$\begin{aligned} D(f_n, g_n) &= E\left((X_0 - \hat{X}_0^{(n)})^2\right) \\ &= E\Big(([X_0 - E(X_0)] - [\hat{X}_0^{(n)} - E(\hat{X}_0^{(n)})] \\ &\qquad + [E(X_0) - E(\hat{X}_0^{(n)})])^2\Big) \\ &= \sigma^2_{X_0} + \sigma^2_{\hat{X}_0^{(n)}} - 2\mathrm{COV}(X_0, \hat{X}_0^{(n)}) \\ &\qquad + [E(X_0) - E(\hat{X}_0^{(n)})]^2 \end{aligned}$$

results in

$$D_X(R) = \lim_{n\to\infty} \left(\sigma^2_{X_0} + (1-2a_n)\sigma^2_{\hat{X}_0^{(n)}} + [E(X_0) - E(\hat{X}_0^{(n)})]^2\right). \quad (31)$$

Subtracting (30) from (31) yields

$$\lim_{n\to\infty} \left((1-a_n)^2\sigma^2_{\hat{X}_0^{(n)}} + [E(X_0) - E(\hat{X}_0^{(n)})]^2\right) = 0. \quad (32)$$

Since both terms in the limit are nonnegative, both must converge to zero since the sum does. Convergence of the rightmost term in the sum proves (20). Provided $D_X(R) < \sigma^2_{X_0}$, which is true if $R > 0$, (30) and (32) together imply that $(a_n - 1)^2/a_n^2$ converges to 0 and hence that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{\mathrm{COV}(X_0, \hat{X}_0^{(n)})}{\sigma^2_{\hat{X}_0^{(n)}}} = 1. \quad (33)$$

This proves (21) and with (31) proves (22) and also that

$$\lim_{n\to\infty} \mathrm{COV}(X_0, \hat{X}_0^{(n)}) = \sigma^2_{X_0} - D_X(R). \quad (34)$$

Finally consider the conditions in terms of the reproduction error. Eq. (23) follows from (20). Eq. (24) follows from (20)–(34) and some algebra. Eq. (25) follows from (23) and the asymptotic optimality of the codes. □

If X is a B-process so that $\Delta_{X|Z}(R) = D_X(R)$, then a similar proof yields corresponding results for the simulation problem. If g_n is an asymptotically optimal (with respect to $\overline{d}_2$ distance) sequence of stationary codes of an IID equiprobable source Z with alphabet B of size $R = \log \|B\|$ which produce a simulated process $\tilde{X}^{(n)}$, then

$$\begin{aligned} \lim_{n\to\infty} E(\tilde{X}_0^{(n)}) &= E(X_0) \\ \lim_{n\to\infty} \sigma^2_{\tilde{X}_0^{(n)}} &= \sigma^2_{X_0} - \Delta_{X|Z}(R). \end{aligned}$$

It is perhaps surprising that when finding the best matching process with constrained rate, the second moments differ.

C. Finite-order distribution Shannon conditions for IID processes

Several code design algorithms, including randomly populating a trellis to mimic the proof of the trellis source encoding theorem [38], are based on the intuition that the guiding principle of designing such a system for an IID source should be to produce a code with marginal reproduction distribution close to a Shannon optimal reproduction distribution [39], [4], [24]. While highly intuitive, we are not aware of any rigorous demonstration to the effect that if a code is optimal or asymptotically optimal, then necessarily its marginal reproduction distribution approaches that of a Shannon optimal. Pearlman [24] was the first to formally conjecture this property of sliding-block codes. The following result addresses this issue. It follows from standard inequalities and Csiszár [2] as summarized in Corollary 1.

Lemma 5: Given a real-valued IID process X with distribution μ_X, assume that f_n, g_n is an asymptotically optimal sequence of stationary source encoder/decoder pairs with common alphabet B of size $R = \log \|B\|$ which produce a reproduction process $\hat{X}^{(n)}$. Then a subsequence of the marginal distribution of the reproduction process, $\mu_{\hat{X}_0^{(n)}}$ converges weakly and in $\mathcal{T}_2$ to a Shannon optimal reproduction distribution. If the Shannon optimal reproduction distribution is unique, then $\mu_{\hat{X}_0^{(n)}}$ converges to it.

Proof: Given the asymptotically optimal sequence of codes, let π_n denote the induced process joint distributions on $(X, \hat{X}^{(n)})$. The encoded process has alphabet size 2^R and hence entropy rate less than or equal to R. Since coding cannot increase entropy rate, the entropy rate of the reproduction (decoded) process is also less than or equal to R. By standard information theoretic inequalities (e.g., [9], p. 193), since the input process is IID we have for all N that

$$\begin{aligned} \frac{1}{N} I(\pi_n^N) &= \frac{1}{N} I(X^N, \hat{X}^N) \geq \frac{1}{N} \sum_{i=0}^{N-1} I(X_i, \hat{X}_i^{(n)}) \\ &= I(X_0, \hat{X}_0^{(n)}) = I(\pi_n^1). \end{aligned} \quad (35)$$

The leftmost term converges to the mutual information rate between the input and reproduction, which is bound above by the entropy rate of the output so that

$$I(X_0, \hat{X}_0^{(n)}) \leq R, \text{ all } n. \quad (36)$$

Since the code sequence is asymptotically optimal, (18) holds. Thus the sequence of joint distributions π_n for $(X_0, \hat{X}_0^{(n)})$ meets the conditions of Corollary 1 and hence $\mu_{\hat{X}_0^{(n)}}$ has a subsequence which converges weakly to a Shannon optimal distribution. If the Shannon optimal distribution μ_{Y_0} is unique, then every subsequence of of $\mu_{\hat{X}_0^{(n)}}$ has a further subsequence which converges to μ_{Y_0}, which implies that $\mu_{\hat{X}_0^{(n)}}$ converges weakly to μ_{Y_0}. The moment conditions (20) and (22)) of Lemma 4 imply that $E[(\hat{X}_0^{(n)})^2]$ converges to $E[(\hat{X}_0)^2]$. The weak convergence of a subsequence of $\mu_{\hat{X}^{(n)}}$ (or the sequence itself) and the convergence of the second moments imply convergence in $\mathcal{T}_2$ [37]. □

Since the source is IID, the N-fold product of a one-dimensional Shannon optimal distribution is an N-dimensional Shannon optimal distribution. If the Shannon optimal marginal distribution is unique, then so is the N-dimensional Shannon optimal distribution. Since Csiszár's [2] results hold for the N-dimensional case, we immediately have the first part of the following corollary.

Corollary 2: Given the assumptions of the lemma, for any positive integer N let $\mu_{\hat{X}^{(n)}}$ denote the N-dimensional joint distribution of the reproduction process $\hat{X}^{(n)}$. Then a subsequence of the N-dimensional reproduction distribution $\mu_{\hat{X}^{(n)}}$ converges weakly and in $\mathcal{T}_2$ to the N-fold product of a Shannon optimal marginal distribution (and hence to an N-dimensional Shannon optimal distribution). If the one dimensional Shannon optimal distribution is unique, then $\mu_{\hat{X}^{(n)}}$ converges weakly and in $\mathcal{T}_2$ to its N-fold product distribution.

Proof: The moment conditions (20) and (22)) of Lemma 4 imply that $E[(\hat{X}_k^{(n)})^2]$ converges to $E[(\hat{X}_k)^2]$ for $k = 0, 1, \ldots, N-1$. The weak convergence of the N-dimensional distribution of a subsequence of $\mu_{\hat{X}^{(n)}}$ (or the sequence

itself) and the convergence of the second moments imply convergence in $\mathcal{T}_2$ [37]. □

There is no counterpart of this result for optimal codes as opposed to asymptotically optimal codes. Consider the Gaussian case where the Shannon optimal distribution is a product Gaussian distribution with variance $\sigma_X^2 - D_X(R)$. If a code were optimal, then for each N the resulting Nth order reproduction distribution would have to equal the Shannon product distribution. But if this were true for all N, the reproduction would have to be the IID process with the Shannon marginals, but that process has infinite entropy rate.

If X is a B-process, then a small variation on the proof yields similar results for the simulation problem: given an IID target source X, the Nth order joint distributions $\mu_{\tilde{X}^{(n)}}$ of an asymptotically optimal sequence of constrained rate simulations $\tilde{X}^{(n)}$ will have a subsequence that converges weakly and in $\mathcal{T}_2$ to an N-dimensional Shannon optimal distribution.

D. Asymptotic uncorrelation

The following theorem proves a result that has often been assumed or claimed to be a property of optimal codes. Define as usual the covariance function of the stationary process $\hat{X}^{(n)}$ by $K_{\hat{X}^{(n)}}(k) = \mathrm{COV}(\hat{X}_i^{(n)}, \hat{X}_{i-k}^{(n)})$ for all integer k.

Theorem 1: Given a real-valued IID process X with distribution μ_X, assume that f_n, g_n is an asymptotically optimal sequence of stationary source encoder/decoder pairs with common alphabet B of size $R = \log \|B\|$ which produce a reproduction process $\hat{X}^{(n)}$. For all $k \neq 0$,

$$\lim_{n\to\infty} K_{\hat{X}^{(n)}}(k) = 0 \tag{37}$$

and hence the reproduction processes are asymptotically uncorrelated.

Proof. If the Shannon optimal distribution is unique, then $\mu_{\hat{X}^{(n)}}$ converges in $\mathcal{T}_2$ to the N-fold product of the Shannon optimal marginal distribution by Corollary 2. As Lemma 6 in the Appendix shows, this implies the convergence of $K_{\hat{X}^{(n)}}(k) = \mathrm{COV}(\hat{X}_k^{(n)}, \hat{X}_0^{(n)})$ to 0 for all $k \neq 0$. □

Taken together these necessary conditions provide straightforward tests for code construction algorithms. Ideally, one would like to prove that a given code construction satisfies these properties, but so far this has only proved possible for the Shannon optimal reproduction distribution property — as exemplified in the next section. The remaining properties, however, can be easily demonstrated numerically for the codes developed next.

IV. An Algorithm for Sliding-Block Simulation and Source Decoder Design

We begin with a sliding-block simulation code which approximately satisfies the Shannon marginal distribution necessary condition for optimality. Matching the code with a Viterbi algorithm encoder then yields a trellis source encoding system.

Consider a sliding-block code g_L of length L of an equiprobable binary IID process Z which produces an output process $\tilde{X}$ defined by

$$\tilde{X}_n = g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1}), \tag{38}$$

where the notation makes sense even if L is infinite, in which case g views a semi-infinite binary sequence. Since the processes are stationary, we emphasize the case $n = 0$. Suppose that the ideal distribution for $\tilde{X}_0$ is given by a CDF F, for example the CDF corresponding to the Shannon optimal marginal reproduction distribution of Lemma 2. Given a CDF F, define the (generalized) inverse CDF F^{-1} as $F^{-1}(u) = \inf\{r\colon F(r) \geq u\}$ for $0 < u < 1$. If U is a uniformly distributed continuous random variable on $(0,1)$, then the random variable $F^{-1}(U)$ has CDF F. The CDF can be approximated by considering the binary L-tuple $u^L = (u_0, u_1, \ldots, u_{L-1})$ comprising the shift register entries as the binary expansion of a number in $(0,1)$:

$$b(u^L) = \sum_{i=0}^{L-1} u_i 2^{-i-1} + 2^{-L-1}, \tag{39}$$

and defining

$$g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1}) = F^{-1}(b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})). \tag{40}$$

If the Z_n is a fair coin flip process, the discrete random variable $b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ is uniformly distributed on the discrete set $\{2^{-L-1}, 2^{-L-1} + 2^{-L}, 2^{-L-1} + 2 \times 2^{-L}, \cdots, 2^{-L-1} + 1 - 2^{-L}\}$, that is, it is a discrete approximation to a uniform $(0,1)$ that improves as L grows, and the distribution of $g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ converges weakly to F, satisfying a necessary condition for an asymptotically optimal sequence of codes. If L is infinite, then the marginal distribution will correspond to the target distribution exactly! This fulfills the necessary condition of weak convergence for an asymptotically optimal code of Lemma 5.

The code as described thus far only provides the correct approximate marginals; it does not provide joint distributions that match the Shannon optimal joint distribution — nor can it exactly since it cannot produce independent pairs. We adapt a heuristic aimed at making pairs of reproduction samples as independent as possible in the sense of $\overline{d}_2$ approximation by modifying the code in a way that decorrelates successive reproductions and hence attempts to satisfy the necessary condition of Theorem 1. Instead of applying the inverse CDF directly to the binary shift register contents, we first permute the binary vectors, that is, the codebook of all 2^L possible shift register contents is permuted by an invertible one-to-one mapping $\mathcal{P} : \{0,1\}^L \to \{0,1\}^L$ and the binary vector $\mathcal{P}(u^L)$ is used to generate the discrete uniform distribution. A randomly chosen permutation $\mathcal{P}$ is used, but *once chosen it is fixed* so that sliding-block decoder is truly stationary. Thus our decoder is

$$g(u^L) = F_{Y_0}^{-1}(b(\mathcal{P}(u^L)), \tag{41}$$

where $F_{Y_0}(y)$ is a Shannon optimal reproduction distribution obtained either analytically (as in the Gaussian case) or from the Rose algorithm (to find the optimum finite support).

Intuitively, the permutation should make the resulting sequence of arguments of the mapping (the number in $(0,1)$ constructed from the permuted binary symbols) resemble an independent sequence and hence cause the sequence of branch

labels to locally appear to be independent. The goal is to satisfy the necessary conditions on joint reproduction distributions of Lemma 2, but we have no proof that the proposed construction has this property. The experimental results to be described show excellent performance approaching the Shannon rate-distortion bound and show that the branch labels are at least uncorrelated, as required by Theorem 1. The permutation is implemented easily by permuting the table entries defining g. For the constrained-rate simulation problem, the permutation does not change the marginal distribution of the coder output, which still converges weakly to the Shannon optimal reproduction distortion as $L \to \infty$, even in the Gaussian case. As in Rose's discussion, this approach discretizes the continuous uniform random variable which is the argument to the mapping (the index space), not the source.

V. Numerical Examples

The random permutation trellis encoder was designed for three common IID test sources: Gaussian, uniform, and Laplacian. The results in terms of both mean squared error (MSE) and signal-to-noise ratio (SNR) are reported for various shift register lengths L indicated by RP_L. The test sequences were all of length 10^6. The results for Gaussian, uniform and Laplacian sources are shown in Table I, II and III respectively.

The distortion-rate function $D_X(R)$ for all three sources are also listed in the tables. For uniform and Laplacian sources, $D_X(R)$ are numerical estimations produced by the Rose algorithm, in both cases, the reported distortions are slightly lower in comparison to the results reported in [17], [21] calculated using the Blahut algorithm [1].

The rate $R = 1$ results of the random permutation trellis coder are compared to previous results of the linear congruential trellis codes (LC) of Eriksson, Anderson, and Goertz [3], trellis coded quantization (TCQ) by Marcellin and Fischer [17], trellis source encoding by Pearlman [24] based on constrained reproduction alphabets and matching the Shannon optimal marginal distribution, a Lloyd-style clustering algorithm conditioned on trellis states by Stewart et al. [32], [33], and the Linde-Gray fake process design [16]. The rate $R = 2$ results are compared with Eriksson et al.'s LC codes and Marcellin and Fisher's TCQ. The rate $R = 3, 4$ results are compared with TCQ which are the only available previous results for these rates.

Eriksson et al.'s LC codes use 512 states for $R = 1$ and 256 states for $R = 2$, which are equivalent to shift register length $L = 10$ and $L = 9$ respectively. Marcellin's TCQ uses 256 states for all rates, corresponding to shift register length 9. Pearlman's results and Stewart's results are for $L = 10$, and Linde/Gray uses a shift register of length 9. The shift register length L is indicated as a subscript for all results.

In the Gaussian example, there are 2^L reproduction levels in the random permutation codes, the result of taking the inverse Shannon optimal CDF, that of a Gaussian zero mean random variable with variance $1 - D_X(R)$, and evaluating it at 2^L numbers in the unit interval. For the uniform, there are 3, 6, 12, and 24 reproduction points for 1, 2, 3, 4 bits chosen by the Rose algorithm for evaluating the first order rate-distortion function. Similarly, for the Laplacian source , there are 9, 17, 31, and 55 reproduction points for 1,2,3,4 bits, respectively.

	Rate(bits)	MSE	SNR(dB)
RP_8	1	0.2989	5.24
RP_9	1	0.2913	5.36
RP_10	1	0.2835	5.47
RP_12	1	0.2740	5.62
RP_16	1	0.2638	5.79
RP_20	1	0.2582	5.88
RP_24	1	0.2557	5.92
RP_28	1	0.2542	5.95
$D_X(R)$	1	0.25	6.02
TCQ9	1	0.3105	5.08
TCQ(opt)_9	1	0.2780	5.56
Pearlman_10	1	0.292	5.35
Stewart_10	1	0.293	5.33
Linde/Gray_9	1	0.31	5.09
LC_10	1	0.2698	5.69
LC(opt)_10	1	0.2673	5.73
	Rate(bits)	MSE	SNR(dB)
RP_24	2	0.0646	11.90
$D_X(R)$	2	0.0625	12.04
TCQ_9	2	0.0873	10.59
TCQ(opt)_9	2	0.0787	11.04
LC(opt)_9	2	0.0690	11.61
RP_24	3	0.0162	17.90
$D_X(R)$	3	0.0156	18.06
TCQ_9	3	0.0237	16.25
TCQ(opt)_9	3	0.0217	16.64
RP_24	4	0.0041	23.92
$D_X(R)$	4	0.0039	24.08
TCQ_9	4	0.0062	22.05

TABLE I
Gaussian Example

Eriksson et al.'s LC codes use 2^{L-1} reproduction points, the Linde/Gray fake process design uses 2^L reproduction points, in both cases, the reproduction points are generated by taking the inverse CDF of the source, evaluating it in the unit interval, and then multiplying with a scaling factor. Stewart also uses 2^L reproduction points, but the reproduction points are obtained through an iterative Lloyd-style training algorithm. Pearlman uses a simpler 4 symbol reproduction alphabet, produced by the Blahut algorithm. Marcellin's TCQ uses 2^{R+1} reconstruction symbols, which are the outputs of the Llyod-Max quantizer. In both LC codes and TCQ, numerical optimization of the reproductions values were used to improve the results. The optimized results for LC codes and TCQ are listed in the tables with the notation "(opt)".

The effectiveness of the random permutation at forcing higher order distributions to look more Gaussian is shown in Fig. 3. The two dimensional scatter plot for adjacent samples with no permutation does not look Gaussian and is clearly highly correlated. When a randomly chosen permutation is used, the plot looks like a 2D Gaussian sample. In both figures, the x and y axis are the value of the samples.

Fig. 4. shows the MSE of the random permutation trellis coder for IID Gaussian at $R = 1$ with various shift register length. The performance has not yet shown to hit a plateau as shift register length increases.

The uniform IID source is of interest because it is simple,

	Rate(bits)	MSE	SNR(dB)
RP_8	1	0.0203	6.13
RP_9	1	0.0195	6.30
RP_10	1	0.0190	6.42
RP_12	1	0.0184	6.55
RP_16	1	0.0179	6.69
RP_20	1	0.0176	6.75
RP_24	1	0.0175	6.78
RP_28	1	0.0174	6.79
$D_X(R)$	1	0.0173	6.84
TCQ_9	1	0.0194	6.33
TCQ(opt)_9	1	0.0183	6.58
LC_10	1	0.0191	6.40
LC(opt)_10	1	0.0179	6.67
	Rate(bits)	MSE	SNR(dB)
RP_24	2	4.02e-03	13.17
$D_X(R)$	2	3.96e-03	13.23
TCQ_9	2	4.24e-03	12.93
TCQ(opt)_9	2	4.18e-03	13.00
LC(opt)_9	2	4.13e-03	13.05
RP_24	3	9.70e-04	19.34
$D_X(R)$	3	9.46e-04	19.45
TCQ_9	3	10.0e-04	19.20
TCQ(opt)_9	3	9.95e-04	19.23
RP_24	4	2.39e-04	25.43
$D_X(R)$	4	2.35e-04	25.50
TCQ_9	4	2.44e-04	25.34

TABLE II
UNIFORM $[0, 1)$ EXAMPLE

	Rate(bits)	MSE	SNR(dB)
RP_8	1	0.2946	5.31
RP_9	1	0.2789	5.55
RP_10	1	0.2671	5.73
RP_12	1	0.2532	5.97
RP_16	1	0.2384	6.23
RP_20	1	0.2306	6.37
RP_24	1	0.2266	6.45
RP_28	1	0.2234	6.51
$D_X(R)$	1	0.2166	6.64
TCQ_9	1	0.3945	4.04
TCQ(opt)_9	1	0.2793	5.54
LC_10	1	0.2529	5.97
LC(opt)_10	1	0.2495	6.03
Pearlman_10	1	0.3058	5.1456
	Rate(bits)	MSE	SNR(dB)
RP_24	2	0.0581	12.36
$D_X(R)$	2	0.0538	12.69
TCQ_9	2	0.1194	9.23
TCQ(opt)_9	2	0.0755	11.22
LC(opt)_9	2	0.0668	11.75
RP_24	3	0.0152	18.18
$D_X(R)$	3	0.0134	18.73
TCQ_9	3	0.0333	14.77
TCQ(opt)_9	3	0.0201	16.96
RP_24	4	0.0046	23.39
$D_X(R)$	4	0.0033	24.79
TCQ_9	4	0.0089	20.53

TABLE III
LAPLACIAN EXAMPLE

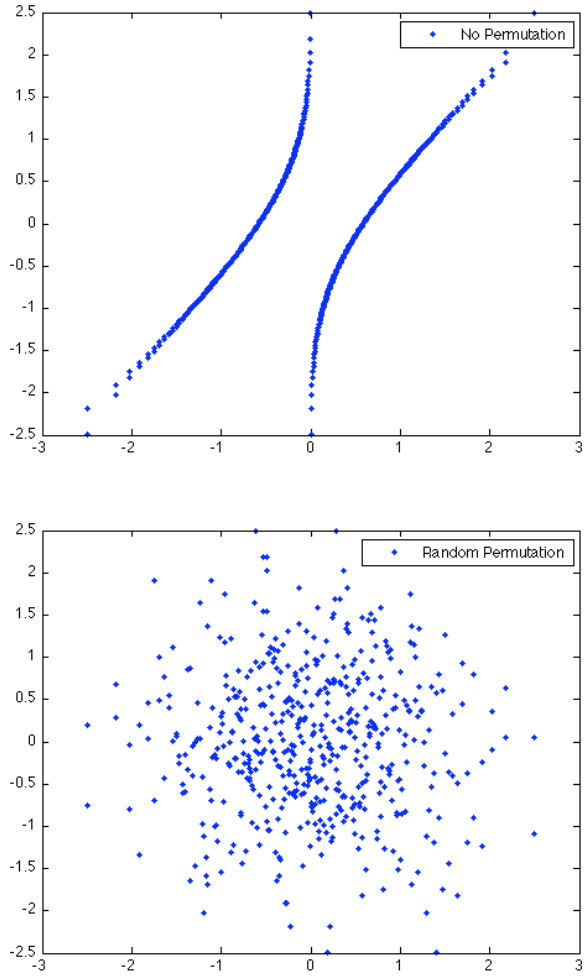

Fig. 3. Scatter plots of fake Gaussian 2-dimensional density: no permutation and random permutation

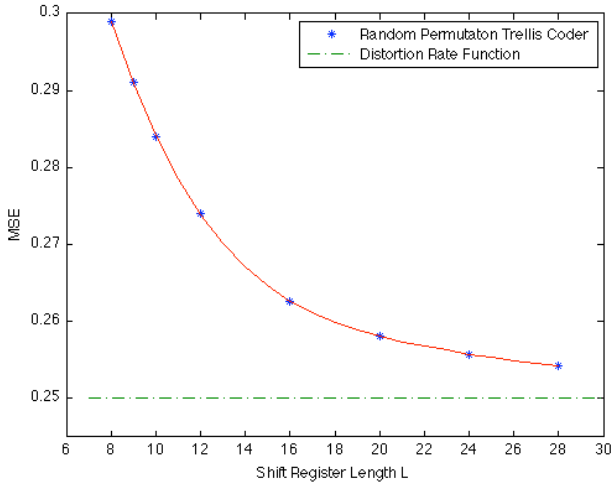

Fig. 4. Performance: 1 bit Gaussian

there is no exact formula for the rate-distortion function with respect to mean-squared error and hence it must be found by numerical means, and because one of the best compression algorithms, trellis-coded quantization (TCQ) is theoretically ideally matched to this example. So the example is an excellent one for demonstrating some of the issues raised here and for comparison with other techniques.

The Rose algorithm yielded a Shannon optimal distribution with an alphabet of size 3 for $R = 1$. The points and their probabilities are shown in Table IV.

y	0.2	0.5	0.8
$p_Y(y)$	0.368	0.264	0.368

TABLE IV
SHANNON OPTIMAL REPRODUCTION DISTRIBUTION FOR THE UNIFORM $(0, 1)$ SOURCE

Plugging the distribution into the random permutation trellis encoder led to a mapping g of (0 0.368) to 0.2, [0.368 0.632] to 0.5, and (0.632 1) to 0.8.

For the Laplacian source of variance 1, the Rose algorithm yielded a Shannon optimal distribution with an alphabet of size 9 for the 1 bit case. The 9 reproduction points and their probabilities are listed in Table V.

y	$\pm$ 4.6273	$\pm$ 3.2828	$\pm$ 2.1654	$\pm$ 1.1063	0
$p_Y(y)$	0.0014	0.0065	0.0285	0.1266	0.6740

TABLE V
SHANNON OPTIMAL REPRODUCTION DISTRIBUTION FOR THE LAPLACIAN SOURCE

For all three test sources — Gaussian, uniform, and Laplacian — the performance of the random permutation trellis source encoder is approaching the Shannon limit. Therefore, it is of interest to estimate the entropy rate of the encoder output bit sequence, which should be close to an IID equiprobable Bernoulli process since an entropy rate near 1 is a necessary condition for approximate optimality [10], [12]. A "plug-in"(or maximum-likelihood) estimator was used for this purpose. The estimator uses the empirical probability of all words of a fixed length in the sequence to estimate the entropy rate. Bit sequences of length 10^6 produced by encoding the Gaussian, uniform, and Laplacian sources with trellis encoder of shift register length $L = 12$ were fed into the estimator, the resulting entropy rate estimation ranges from 0.9993 to 0.9995. For comparison, the estimator yielded entropy rate of 0.9998 for a randomly generated bit sequence of the same length.

Eriksson et al.'s LC results for 1 bit at 512 states (equivalent to shift register length 10) for Gaussian source is better than the random permutation results for the same shift register length. This is likely the result of their exhaustive search over all possible ways of labeling the branches within the constraint of their axioms. A similar approach to the random permutation code would be to search for the permutation that produced the best results. Our results are from randomly chosen permutations, so they reflect performance of the ensemble average (which we believe may eventually lead to a source coding theorem using random coding ideas). All permutations have the same marginals, but some permutations will have better higher order distributions. Such an optimization is feasible only for small L. We tested an optimization by exhaustion for $L = 3$ and found that the best MSE (SNR) was 0.3262 (4.8647), while the average MSE (SNR) for all permutations was 0.3852 (4.1431). This demonstrates that the best permutation can provide notable improvement over the average, but we have no efficient search algorithm for finding optimum permutations.

APPENDIX A
PROOF OF LEMMA 3

The encoded and decoded processes are both stationary and ergodic since the original source is. From (8) and the source coding theorem,

$$\begin{aligned} D(f_n, g_n) &= E[d(X_0, X_0^{(n)})] \geq \overline{d}(\mu_X, \mu_{\hat{X}^{(n)}}) \\ &\geq \inf_{\nu: H(\nu) \leq R} \overline{d}(\mu_X, \nu) = D_X(R). \end{aligned}$$

Here first inequality follows since stationary coding reduces entropy rate, and so $R \geq H(U^{(n)}) \geq H(\hat{X}^{(n)})$. Since the leftmost term converges to the rightmost, the first equality of the lemma is proved.

Standard inequalities of information theory yield

$$R \geq H(\hat{X}^{(n)}) \geq I(X, \hat{X}^{(n)}) \geq R_X(D(f_n, g_n)).$$

where second inequality follows since mutual information rate is bounded above by entropy rate, and the third inequality follows from the process definition of the Shannon rate-distortion function [18], [14]. Taking the limit as $n \to \infty$, the rightmost term converges to R since the code sequence is asymptotically optimal and the Shannon rate-distortion function is a continuous function of its argument (except possibly at $D = 0$). Thus $\lim_{n\to\infty} H(U^{(n)}) = \lim_{n\to\infty} H(\hat{X}^{(n)}) = R$. proving the second equality of the lemma.

The final part requires Marton's inequality [19] relating Ornstein's $\overline{d}$ distance and relative entropy when one of the processes is IID. Suppose that μ_U and μ_Z are stationary process distributions for two processes with a common discrete alphabet and that μ_{U^N} and μ_{Z^N} denote the finite dimensional distributions. For any integer N the relative entropy or informational divergence is defined by

$$H(\mu_{U^N} \| \mu_{Z^N}) = \sum_{u^N} \mu_{U^N}(u^N) \log \frac{\mu_{U^N}(u^N)}{\mu_{Z^N}(u^N)}.$$

In our notation Marton's inequality states that if U is a stationary ergodic process and Z is an IID process, then

$$N^{-1}\mathcal{T}_0(\mu_{U^N}, \mu_{Z^N}) \leq \left[\frac{\ln 2}{2N} H(\mu_{U^N} \| \mu_{Z^N}))\right]^{1/2}.$$

Since Z is an IID equiprobable process with alphabet size 2^R,

$$N^{-1}\mathcal{T}_0(\mu_{U^N}, \mu_{Z^N}) \leq \left[\frac{\ln 2}{2N}(NR - H(U^N))\right]^{1/2}$$

and taking the limit as $N \to \infty$ yields (in view of property (3) of the $\overline{d}$ distance)

$$\overline{d}_0(\mu_U, \mu_Z) \le \left[\frac{\ln 2}{2}(R - H(U))\right]^{1/2}.$$

Applying this to $U^{(n)}$ and taking the limit using the previous part of the lemma completes the proof. □

Lemma 6: Let μ^N denote the N-fold product of a probability distribution μ on the real line such that $\int x^2 d\mu(x) < \infty$. Assume $\{\nu_n\}$ is a sequence of probability distribution on $\mathbb{R}^N$ such that $\lim_{n\to\infty} \mathcal{T}_2(\mu^N, \nu_n) = 0$. If $Y_1^{(n)}, Y_2^{(n)}, \ldots, Y_N^{(n)}$ are random variables with joint distribution ν_n, then for all $i \ne j$,

$$\lim_{n\to\infty} E\big[\big(Y_i^{(n)} - E(Y_i^{(n)})\big)\big(Y_j^{(n)} - E(Y_j^{(n)})\big)\big] = 0.$$

Proof. The convergence of ν_n to μ^N in $\mathcal{T}_2$ distance implies that there exist IID random variables $Y_1, \ldots, Y_N$ with common distribution μ and a sequence or N random variables $Y_1^{(n)}, Y_2^{(n)}, \ldots, Y_N^{(n)}$ with joint distribution ν_n, all defined on the same probability space, such that

$$\lim_{n\to\infty} E[(Y_i^{(n)} - Y_i)^2] = 0, \quad i = 1, \ldots, N. \tag{42}$$

First note that this implies for all i

$$\lim_{n\to\infty} E[(Y_i^{(n)})^2] = E[Y_i^2]. \tag{43}$$

Also, $\lim_{n\to\infty} E|Y_i^{(n)} - Y_i| = 0$ (Cauchy-Schwarz), so that for all i,

$$\lim_{n\to\infty} E(Y_i^{(n)}) = E(Y_i). \tag{44}$$

Now the statement is direct convergence of the fact that in any inner product space, the inner product is jointly continuous. To be more concrete, letting $\langle X, Y\rangle = E(XY)$ and $\|X\| = [E(X^2)]^{1/2}$ for random variables X and Y with finite second moment defined on this probability space, we have the bound

$$\begin{aligned}
&\big|\langle Y_i^{(n)}, Y_j^{(n)}\rangle - \langle Y_i, Y_j\rangle\big| \\
&\quad\le \big|\langle Y_i^{(n)}, Y_j^{(n)} - Y_j\rangle\big| + \big|\langle Y_i^{(n)} - Y_i, Y_j\rangle\big| \\
&\quad\le \|Y_i^{(n)}\|\,\|Y_j^{(n)} - Y_j\| + \|Y_i^{(n)} - Y_i\|\,\|Y_j\|.
\end{aligned}$$

Since $\|Y_i^{(n)}\|$ converges to $\|Y_i\|$ by (43) and $\|Y_i^{(n)} - Y_i\|$ converges to zero by (42), we obtain that $\langle Y_i^{(n)}, Y_j^{(n)}\rangle$ converges to $\langle Y_i, Y_j\rangle$, i.e,

$$\lim_{n\to\infty} E(Y_i^{(n)} Y_i^{(n)}) = E(Y_i Y_j) = E(Y_i)E(Y_j)$$

since Y_i and Y_j are independent if $i \ne j$. This and (44) imply the lemma statement. □

REFERENCES

[1] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460-473, Jul. 1972.

[2] I. Csiszár, "On an extremum problem in information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1–2, pp. 57–71, 1974.

[3] T. Eriksson, J.B. Anderson, and N. Goertz, " Linear congruential trellis source codes: design and analysis," *IEEE Transactions on Communications*, vol. 55, no. 9, pp. 1693–1701, Sept. 2007.

[4] W. A. Finamore and W. A. Pearlman. "Optimal encoding of discrete-time, continuous-amplitude, memoryless sources with finite output alphabets," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 144–155, Mar. 1980.

[5] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, 1968.

[6] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, 1992.

[7] R. M. Gray, "Sliding–block source coding," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 4, pp. 357–368, Jul. 1975.

[8] R. M. Gray, "Time–invariant trellis encoding of ergodic discrete–time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–23, pp. 71–83, Jan. 1977.

[9] R. M. Gray, *Entropy and Information Theory*, Springer–Verlag, New York, 1990.

[10] R. M. Gray, "Source coding and simulation," *IEEE Information Theory Society Newsletter*, vol. 58, no. 4, Dec. 2008.

[11] R. M. Gray, *Probability, Random Processes, and Ergodic Properties: Second Edition*, Springer, New York, 2009.

[12] R. M. Gray and T. Linder, "Bits in asymptotically optimal lossy source codes are asymptotically Bernoulli," *Proc. 2009 Data Compression Conference (DCC)*, pp. 53–62, Mar. 2008.

[13] R. M. Gray, D. L. Neuhoff and P. C. Shields, "A generalization of Ornstein's d–bar distance with applications to information theory," *Annals of Probability*, vol. 3, no. 2, pp. 315–328, Apr. 1975.

[14] R. M. Gray, D. L. Neuhoff and D. S. Ornstein, "Nonblock source coding with a fidelity criterion," *Annals of Probability*, vol. 3, no. 3, pp. 478–491, Jun. 1975.

[15] L. V. Kantorovich, "On a problem of Monge," *Dokl. Akad. Nauk*, vol. 3, 225–226, 1948.

[16] Y. Linde and R. M. Gray, "A fake process approach to data compression," *IEEE Trans. Commun.*, vol. COM–26, pp. 840–847, Jun. 1978.

[17] M. Marcellin and T. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 92–93, Jan. 1990.

[18] K. Marton, "On the rate distortion function of stationary sources," *Probl. Contr. Inform. Theory*, vol. 4, pp. 289-297, 1975.

[19] K. Marton, "Bounding $\bar{d}$ distance by informational divergence: a method to prove measure concentration," *Annals of Probability*, vol. 24, pp. 857–866, 1997.

[20] G. Monge, "Mémoire sur la théorie des déblais et des remblais," in *Histoire de lAcadémie Royale des Sciences de Paris*, pp. 666–704, 1781.

[21] P. Noll and R. Zelinski, "Bounds on quantizer performance in the low bit-rate region," *IEEE Trans. Commun.*, vol. COM-26, no. 2, pp. 300-304, Feb. 1978.

[22] D. Ornstein, "An application of ergodic theory to probability theory," *Annals of Probability*, vol. 1, no. 1, pp. 43–58, 1973.

[23] D. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*, Yale University Press, New Haven, 1975.

[24] W. A. Pearlman, "Sliding-block and random source coding with constrained size reproduction alphabets," *IEEE Trans. Commun.*, vol. COM-30, no. 8, pp. 1859-1867, Aug. 1982.

[25] M. S. Pinsker, "Information and information stability of random variables and processes," Holden Day, San Francisco, 1964.

[26] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems Vol. I: Theory, Vol. II: Applications*. Probability and its applications. Springer-Verlag, New York, 1998.

[27] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1939–1952, Nov. 1994.

[28] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record, Part 4*, pp. 142–163, 1959.

[29] P. C. Shields, "Stationary coding of processes," *IEEE Trans. Inform. Theory*, vol. 25, no. 3, pp. 283–291, May 1979.

[30] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, AMS Graduate Studies in Mathematics, Vol. 13, American Mathematical Society, 1996.

[31] M. Smorodinsky, "A partition on a Bernoulli shift which is not weak Bernoulli," *Math. Systems Theory*, vol. 5, no. 3, pp. 201–203, 1971.

[32] L. Stewart, *Trellis data compression*, Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Stanford, CA 94305, June 1981.

[33] L. Stewart, R. M. Gray, and Y. Linde, "The design of trellis waveform coders," *IEEE Trans. Commun.*, vol. COM-30, no. 4, pp. 702-710, Apr. 1982.

[34] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86. Jan. 1996.

[35] R. J. van der Vleuten and J. H. Weber, "Construction and evaluation of trellis-coded quantizers for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 853–87 , May 1995.
[36] L. N. Vasershtein. "Markov processes on countable product space describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, pp. 64–73, 1969.
[37] C. Villani, *Optimal Transport, Old and New*, Grundlehren der mathematischen Wissenschaften, vol. 338, Springer 2009.
[38] A. J. Viterbi and J. K. Omura, "Trellis Encoding of Memoryless Discrete-Time Sources with a Fidelity Criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 3, May 1974.
[39] S. G. Wilson and D. W. Lytle, "Trellis coding of continuous-amplitude memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 404–409, May 1977.