2010 Data Compression Conference

Stationary and Trellis Encoding for IID Sources and Simulation [1]

Mark Z. Mao Robert M. Gray
Dept of Electrical Engineering, Stanford University, Stanford, CA 94305
markmao@stanford.edu rmgray@stanford.edu

Abstract

Necessary conditions for asymptotically optimal sliding-block or stationary codes for source coding and rate-constrained simulation are presented and applied to a design technique for trellis-encoded source coding and rate constrained simulation of memoryless sources.

1 Introduction

One view of the basic goal of Shannon source coding with a fidelity criterion or lossy data compression is to convert 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. 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 constrained entropy rate? For example, suppose the simulator or model operates on fair coin flips and produces one approximately Gaussian sample for each flip. How good an approximation can be generated in this way, where *good* can be quantified by using a metric on random processes such as the Monge/Kantorovich/Wasserstein/Ornstein distance [12, 21, 16, 6]?

It has been known for over three decades that the two problems of source coding and simulation are intimately related, as is obvious from the general block diagram of Figure 1, in which the right half of the source coding system 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 [11]), if the source code operates near the Shannon optimum, then the bits being transmitted should be close to an IID equiprobable Bernoulli process. This is an example of a necessary condition for a sequence of codes

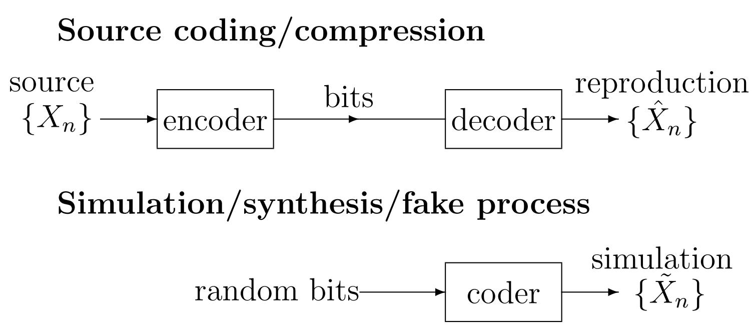

Figure 1: Source coding and simulation

[1]This work was supported by the National Science Foundation under Grant CCF-0846199-000.

1068-0314/10 $26.00 © 2010 IEEE
DOI 10.1109/DCC.2010.8

to be asymptotically optimal. 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 where the codes are stationary, that is, the mapping from input sequences to output sequences is shift-invariant (shifting the input results in a corresponding shift of the output), and the source $\{X_n\}$ is a B-process, that is, a stationary coding of an IID process [16]. For example, regular Gaussian processes are B-processes. Related results for asymptotically long block codes were developed Steinberg and Verdu in 1996 [20].

We describe new necessary conditions for a sequence of stationary codes to be asymptotically optimal as source codes or simulation codes and the conditions motivate a new design technique for trellis source encoding. Since this paper was submitted to DCC, joint work with T. Linder has led to extensions of the necessary conditions reported here. The extensions along with supporting details for the results described here may be found in [14].

2 Preliminaries

Stationary codes A stationary code can be viewed as a possibly nonlinear time-invariant filter. 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. 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 and 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.

Fidelity and distortion A fidelity criterion for a source coding system consists of a family of distortion measures (or loss functions or cost functions) $d_N(x^N, y^N)$, $N = 1, 2, \ldots$ mapping N tuples $x^N = (x_0, x_1, \ldots, x_{N-1}) \in A^N$, where A^N is the N-fold Cartesian product of a source alphabet A such as the real line, and $y^N \in \hat{A}^N$, where $\hat{A}$ is a reproduction alphabet, into the nonnegative real line.

Given random vectors X^N, Y^N with a joint distribution π^N, the average distortion is defined by the expectation $d_N(\pi^N) = E_{\pi^N}[d_N(X^N, Y^N)]$. We assume a *single-letter* distortion measure defined in terms of a distortion measure $d(x, y)$, $x \in A, y \in \hat{A}$, by $d_N(x^N, y^N) = \sum_{i=0}^{N-1} d(x_i, y_i)$. Given a stationary pair process $\{X_n, Y_n\}$, $N^{-1}E[d_N(X^N, Y^N)] = N^{-1}d_N(\pi^N) = E[d(X_0, Y_0)] = d_1(\pi^1)$ and hence a measure of the fidelity of a code (f, g) coding X_n into a reproduction $\hat{X}_n$ is $D(f, g) = E[d(X_0, \hat{X}_0)]$. We emphasize the common squared error distortion, $d(x, y) = (x - y)^2$

and the Hamming distortion, $d(x, y) = 0$ if $x = y$ and 1 otherwise.
Optimal source coding The operational distortion-rate function $\delta_X(R)$ is the infimum of $D(f, g)$ over all stationary encoders f and decoders g with common encoder output/decoder input alphabet $B = \{0, 1\}$. The extension to alphabets of size 2^R for integer R is straightforward and some results are given in this generality.
Process distance measures A distortion measure d induces a natural notion of a "distance" between random processes. The basic idea is an extension to general distortion measures of Ornstein's $\overline{d}$ (d-bar) distance (which is based on the the Hamming distortion) or, equivalently, the extension of the Monge/Kantorovich/Wasserstein optimal transportation cost or distance from random vectors (see, e.g., Villani [21]) to random processes. 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 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. Any π having the prescribed marginals μ_X and μ_Y is called a *coupling* of μ_X and μ_Y. The most important special case is when the cost function is a nonnegative power of an underlying metric, say $d(x, y) = m(x, y)^r$, 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 itself 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 cases of squared error and Hamming distance, respectively.

Given two stationary processes with process distributions μ_X and μ_X, let μ_{X^N} and μ_{Y^N} denote the induced N-dimensional distributions. Let d_N be a single-letter fidelity criterion with $d_1 = d$. Define the process distance

$$\overline{d}(\mu_X, \mu_Y) = \sup_N N^{-1} \mathcal{T}(\mu_{X^N}, \mu_{Y^N}) = \inf_{\pi \in \mathcal{P}(\mu_X, \mu_Y)} E_\pi[d(X_0, Y_0)].$$

Many properties of the $\overline{d}$ and generalized $\overline{d}$ (or $\overline{\rho}$ distance) are detailed in [16, 6, 10].
Optimal rate-constrained simulation The $\overline{d}$ distance can be used to describe both the optimal source coding and the optimal simulation. 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 (here usually assumed to be fair coin flips). The optimal simulation of the process $X = \{X_n\}$ given the process $Z = \{Z_n\}$ is the solution (if it exists) to$\Delta_{X|Z}(R) = \inf_{f \in \mathcal{C}(B, \hat{A})} \overline{d}(\mu_X, \mu_{\bar{f}(Z)})$, where $\mu_{\bar{f}(Z)} = \mu_Z \bar{f}^{-1}$ is the process distribution resulting from a stationary coding of Z using f. An alternative definition can be stated in terms of the Shannon entropy rate of a random process. Define as usual the Shannon entropy and entropy rate by

$$\begin{aligned} H(X^N) &= H(\mu_{X^N}) = \begin{cases} -\sum_{x^N} \mu_{X^N}(x^N) \log \mu_{X^N}(x^N) & A_X \text{ discrete} \\ \infty & \text{otherwise} \end{cases} \\ H(X) &= H(\mu_X) = \inf_N H(X^N)/N = \lim_{N \to \infty} H(X^N)/N. \end{aligned}$$

If the source is stationary and ergodic, then [8]

$$\Delta_{X|Z}(R) = \inf_{B-\text{processes } \nu : H(\nu) \le R} \overline{d}(\mu_X, \nu). \tag{1}$$

An characterization of the operational distortion rate function which resembles strongly the optimal simulation description is [7]

$$\delta_X(R) = \inf_{\nu: H(\nu) \le R} \overline{d}(\mu_X, \nu) \le \Delta_{X|Z}(R), \tag{2}$$

where the infimum is over all stationary and ergodic processes. If the source X is also a B-process, then$\Delta_{X|Z}(R) = \delta_X(R)$.

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 average mutual information over all possible 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 mutual information rate (if it exists) is then $I(X;Y) = \lim_{N\to\infty} N^{-1} I(X^N, Y^N)$. The Shannon rate-distortion function is defined by

$$\begin{aligned} R_X(D) &= \inf_N N^{-1} R_{X^N}(D) = \lim_{N\to\infty} N^{-1} R_{X^N}(D) \\ R_{X^N}(D) &= \inf_{\pi^N: \pi^N \in \mathcal{P}(\mu_{X^N}), N^{-1} E d_N(\pi^N) \le D} N^{-1} I(\pi^N) \end{aligned}$$

where the infimum is over all joint distributions π^N for X^N, Y^N with marginal distribution μ_{X^N} ($\pi^N \in \mathcal{P}(\mu_{X^N})$) and $N^{-1}E[d(X^N, Y^N)] \le R$. The dual function, Shannon's distortion-rate function $D_X(R)$, is defined similarly. Source coding theorems show that under suitable conditions $\delta_X(R) = D_X(R)$ for sliding-block codes as well as for block codes [7, 9].

If the process X is IID, then

$$R_X(D) = R_{X_0}(D) = \inf_{\pi: \pi \in \mathcal{P}(\mu_{X_0}), Ed(X_0, Y_0) \le D} I(X_0, Y_0) \tag{3}$$

and much can be said about evaluating $R_X(D)$. The following lemma is a rephrasing of Csiszár's generalization [2] of the Gallager's [4] Kuhn-Tucker optimization. It is a combination of Csiszár's Lemma 1.2, his corollary to Lemma 1.3, and his equations (1.11) and (1.15).

Lemma 1 *If $R_{X_0}(D) < \infty$, then $R_{X_0}(D) = \max_{\beta \ge 0} (\Gamma(\beta) - \beta D)$, where*

$$\Gamma(\beta) = \inf_{\pi \in \mathcal{P}(\mu_{X_0})} (I(\pi) + \beta D(\pi)) = \inf_{\mu_{Y_0}} \int d\mu_{X_0}(x) \log \frac{1}{\int d\mu_{Y_0}(y) e^{-\beta d(x,y)}}. \tag{4}$$

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 (4), then $D = d(\pi)$ and $R(D) = I(\pi)$.*

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_0} on the reproduction alphabet. If a minimizing π exists, then the resulting marginal distribution for μ_{Y_0} is called a *Shannon optimal reproduction distribution*.

Csiszár [2]'s Theorem 2.2, its proof, and the extension of the reproduction space from compact to Polish discussed at the bottom of p. 66 of [2] yields

Lemma 2 *Given a random variable X_0 with an alphabet A, a Euclidean space with Euclidean metric $\|x-y\|$, 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 (3). Suppose that $\pi^{(n)}$, $n=1,2,\ldots$ is sequence of distributions on $A\times\hat{A}$ with marginals μ_{X_0} and $\mu_{Y_0}^{(n)}$ for which*

$$\begin{aligned} I(\pi^{(n)}) &= I(X_0, Y_0^{(n)}) \leq R; n=1,2,\ldots && (5)\\ \lim_{n\to\infty} E[d(X_0, Y_0^{(n)})] &= D_{X_0}(R). && (6)\end{aligned}$$

Then $\mu_{Y_0}^{(n)}$ converges weakly to a Shannon optimal reproduction distribution.

Rose [18] argues that the minimization of (4) over all distributions on the reproduction alphabet can be replaced by$\Gamma(\ \beta)=\inf_f[-\int d\mu_{X_0}(x)\log\int_0^1 d\lambda(u)e^{-\beta d(x,f(u))}]$, a minimization over all measurable mappings from the unit interval into the reproduction alphabet, where λ denotes the the uniform distribution on $[0,1)$. He provides an alternative to the traditional optimization of Lemma 1 using the Blahut algorithm [1] by an optimization over mappings using a form of annealing. A significant advantage of the Rose approach is that it allows one to avoid the discretization of the input alphabet of a continuous random variable and instead shifts the focus to the reproduction alphabet. Rose argues that for a continuous alphabet source, the Shannon optimal reproduction distribution will have continuous or infinite support only in the special case where the Shannon lower bound to the rate distortion function holds with equality, e.g., in the case of an IID Gaussian source and squared error. In other cases, the Shannon optimal reproduction distribution will have finite support.

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 Fig. 2. Here we conform to traditional notation and reverse the time order inside the function so that the most recent symbol is leftmost. 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 "populated" by the decoder, also shown in the figure. The Viterbi algorithm yields a block encoder matched to a 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 (see, e.g., [9]).

The original source coding theorem and most of the subsequent literature treat time-varying trellis codes [22, 23, 17] which attempt to randomly populate each level of the trellis with independent random variables generated by a Shannon optimal reproduction distribution. Time-invariant systems are considered in [7, 8, 13, 19, 3, 15], and of these the trellis-coded quantization (TCQ) system of Marcellin and Fischer [15] has provided both the best performance and low complexity. A discussion and comparison of the various approaches may be found in the full version of the paper[14].

3 Asymptotically Optimal Codes

A sequence of sliding-block codes f_n, g_n, $n=1,2,\ldots$, for source coding is *asymptotically optimal* if $\lim_{n\to\infty} D(f_n,g_n)=\delta_X(R)=D_X(R)$. An optimal code is trivially

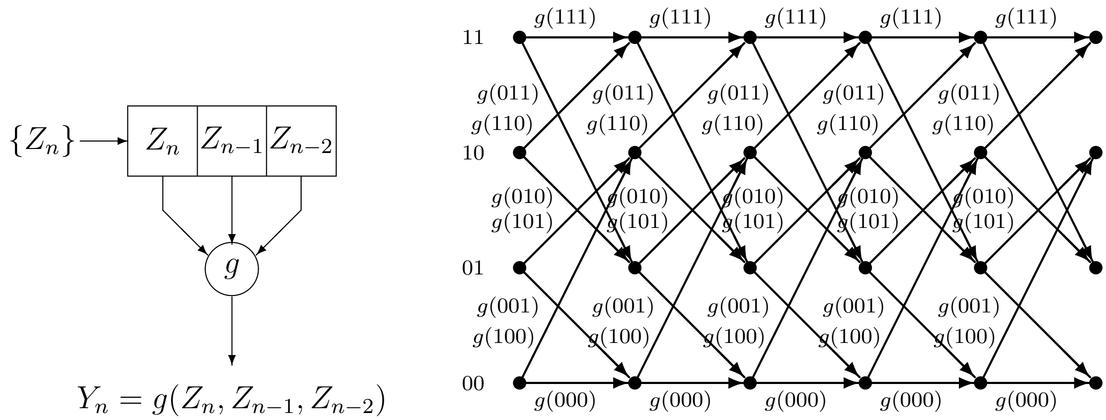

Figure 2: Simple sliding-block code with trellis

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 sequence of simulation codes g_n is asymptotically optimal if $\lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\bar{g}_n(Z)}) = \Delta(X|Z)$.

Process necessary conditions

The source coding theorem for stationary codes and [11] imply the following conditions. Given a real-valued stationary ergodic process X, suppose that f_n, g_n is an asymptotically optimal sequence of stationary source codes for X which yield reproductions $\hat{X}^{(n)}$ and the encoder output process is $U^{(n)}$, then $\lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\hat{X}^{(n)}}) = D_X(R)$ and $\lim_{n\to\infty} \overline{d}(\mu_Z, \mu_{U^{(n)}}) = 0$, where Z is an equiprobable IID source with alphabet of size 2^R (fair coin flips when $R = 1$).

Moment necessary conditions

The first result concerns the squared error distortion and resembles a standard result for scalar and vector quantizers (see, e.g., [5], Lemmas 6.2.2 and 11.2.2).

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 $\hat{X}^{(n)}$, then

$$\lim_{n\to\infty} E(\hat{X}_0^{(n)}) = E(X_0), \lim_{n\to\infty} \frac{\text{COV}(X_0, \hat{X}_0^{(n)})}{\sigma^2_{\hat{X}_0^{(n)}}} = 1, \lim_{n\to\infty} \sigma^2_{\hat{X}_0^{(n)}} = \sigma^2_{X_0} - D_X(R).$$

Comment: Similar to the conditions for optimal scalar quantizers, necessary conditions for a sequence of codes to be asymptotically optimal are that in the limit: **1.** the means of the source and the reproduction are equal, **2.** the reproduction error $e_n = \hat{X}_n - X_n$ is uncorrelated with the reproduction, and **3.** the variance of the source equals the sum of the expected squared error and the reproduction variance.

The details of the proofs can be found in [14], but the idea is simple. The performance using a decoder g_n can be no smaller than the minimum achievable average distortion using all decoders of the form $ag_n + b$. Finding the best a, b is a simple optimal affine prediction problem and the minimum mean squared error can be described in terms of the second order moments. This improved performance, however, is achieved by a code which still has entropy rate less than R since the affine operation can not increase entropy. The performance is still bound below by the distortion-rate function and sandwiching the performance between an asymptotically optimal code sequence and the distortion-rate function yields all of the conditions. A similar proof yields the corresponding result for the simulation problem.

Given a real-valued stationary ergodic process X with distribution μ_X, assume that 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

$$\lim_{n\to\infty} E(\tilde{X}_0^{(n)}) = E(X_0), \lim_{n\to\infty} \sigma^2_{\tilde{X}_0^{(n)}} = \sigma^2_{X_0} - D_X(R).$$

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

Marginal distribution Shannon condition for IID processes

Several code design algorithms, including randomly populating a trellis to mimic the proof of the trellis source encoding theorem [22], are based on the intuition that the guiding principal 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 [23, 3, 17]. The following results back up this intuition. They follow from standard inequalities and Csiszár [2] as summarized in Lemma 2. Details may be found in [14].

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 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.

Given a real-valued IID process X with distribution μ_X, assume that g_n, $n = 1, 2, \ldots$ is an asymptotically optimal sequence of sliding-block codes of an IID process Z with alphabet B of size $R = \log \|B\|$ which produce simulation processes $\tilde{X}^{(n)}$ for which $\lim_{n\to\infty} \overline{d}_2(\mu_X, \mu_{\tilde{X}^{(n)}}) = D_X(R)$. Then the marginal distribution of $\mu_{\tilde{X}_0^{(n)}}$ converges weakly and in $\mathcal{T}_2$ to a Shannon optimal reproduction distribution.

This result is extended to finite-dimension distortion-rate functions in [14].

4 An algorithm for sliding-block decoder design

Consider a sliding-block code g of length L of an equiprobable binary IID process Z which produces an output process $\tilde{X}$ defined by $\tilde{X}_n = g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$. Since the processes are stationary, consider $n = 0$. Suppose that the ideal distribution for $\tilde{X}_0$ is given by a CDF F corresponding to the Shannon optimal marginal reproduction distribution of Lemma 2. If U is a uniformly distributed continuous random variable on $[0, 1)$, then the random variable $F^{-1}(U)$, where F^{-1} is the (generalized) inverse CDF, 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 $[0, 1)$: $b(u^L) = \sum_{i=0}^{L-1} u_i 2^{-i-1} + 2^{-L-1}$, and defining $g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ by $F^{-1}(b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1}))$. The random variable

$b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ is uniformly distributed on a discrete set if Z is a fair coin flip process, and, as L grows, converges weakly to a uniform [01) random variable, hence $g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ converges to F in $\overline{d}_2$, 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! If the Shannon optimal distribution is continuous, the continuous CDF is used. Otherwise the Rose algorithm [18] is used to directly discretize the reproduction alphabet. The inverse CDF can supply approximately the correct marginal distribution, but successive outputs will be highly dependent. Hence 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 is used, but *once chosen it is fixed* so that sliding-block decoder and its trellis are time-invariant. The resulting decoder is $g(u^L) = F_{Y_0}^{-1}(b(\mathcal{P}(u^L)))$, 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).

5 Numerical Examples

The random permutation trellis encoder was designed for Gaussian and uniform sources. The results in terms of both mean squared error (MSE) and signal-to-noise ratio (SNR in dB) are reported for various shift register lengths L indicated by RPL. The results are compared with the TCQ of Marcellin and Fischer [15], and the Gaussian results are compared with a variety of techniques [17, 19, 13]. The test sequences were all of length 10^6. The results are shown in Figure 3. In the Gaussian example, there are 2^L reproduction levels, 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 source, there are 3 reproduction points chosen by the Rose algorithm for evaluating the first order rate-distortion function. The number of states was 2^{L-1} in each example. Pearlman's results and Stewart's results are for $L = 10$, but Pearlman uses

	MSE	SNR
RP8	0.2989	5.24
RP9	0.2913	5.36
RP10	0.2835	5.47
RP12	0.2740	5.62
RP16	0.2638	5.79
RP20	0.2582	5.88
RP24	0.2557	5.92
RP28	0.2542	5.95
$D_X(R)$	0.25	6.02
TCQ	0.2780	5.56
Pearlman	0.292	5.35
Stewart	0.293	5.33
Linde/Gray	0.31	5.09

	MSE	SNR
RP8	0.0203	6.13
RP9	0.0195	6.30
RP10	0.0190	6.42
RP12	0.0184	6.55
RP16	0.0179	6.69
RP20	0.0176	6.75
RP24	0.0175	6.78
RP28	0.0174	6.79
$D_X(R)$	0.0173	6.84
TCQ	0.0183	6.58

Figure 3: $R = 1$: Gaussian (left). Uniform $[0, 1)$ (right)

the simpler 4 symbol reproduction alphabet. The Linde/Gray and Marcellin/Fischer codes use a shift register of length 9, but the latter uses 2^{R+1} reconstruction symbols. Figure 4(left) shows the MSE of the random permutation trellis coder for IID Gaussian at $R = 1$ with various shift register lengths.

The Gaussian source is of particular interest because of the challenge of simulating a good Gaussian imitation with only one bit per symbol. Figure 4(right) shows visually the resulting imitation together with a sample of a real Gaussian (to the accuracy of a digital computer) and the simple one-bit scalar quantized Gaussian. The picture shows the result of source coding the true Gaussian into the best fake available on the trellis.

The uniform IID source is of interest because it is simple, 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 are 0.2, 0.5, 0.8 and the probabilities are 0.368, 0.264, 0.368 respectively.

For both test sources the performance of the random permutation trellis source encoder approaches the Shannon limit. A "plug-in" (or maximum-likelihood) estimator was used to estimate the entropy rate of the channel bits. In all cases the estimate exceeded .999. The other necessary conditions were satisfied within similar numerical accuracy. More results at rates $R = 2, 3, 4$, results for the Laplacian source, and comparisons with other approaches can be found in the full paper[14].

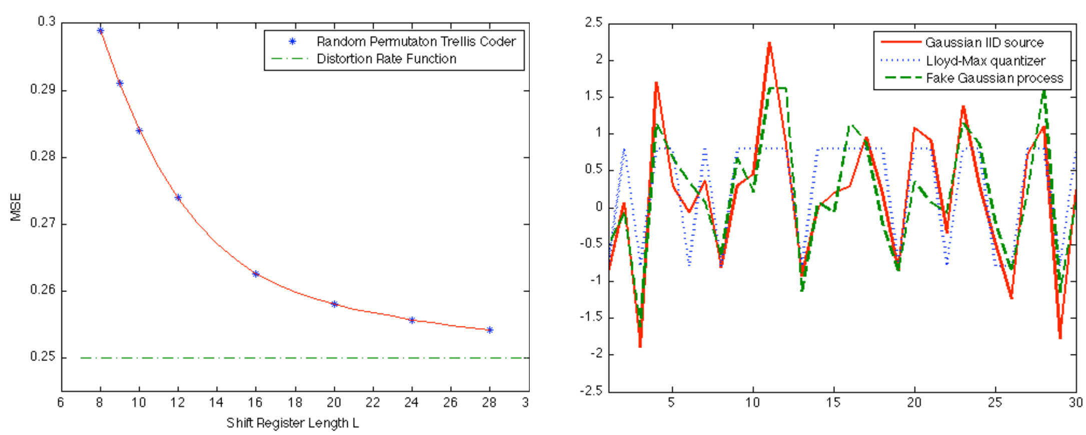

Figure 4: Left: 1 bit Gaussian performance. Right: 1 bit fake Gaussian

References

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

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

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

[6] 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, April 1975.

[7] 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, June 1975.

[8] R. M. Gray, “Time–invariant trellis encoding of ergodic discrete–time sources with a fidelity criterion,” *IEEE Trans. on Info. 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, *Probability, Random Processes, and Ergodic Properties: Second Edition*, August 2009, Springer, New York.

[11] R. M. Gray and T. Linder, “Bits in Asymptotically Optimal Lossy Source Codes are Asymptotically Bernoulli,” *Proceedings 2009 Data Compression Conference (DCC)*, March 2008, pp. 53–62.

[12] L. V. Kantorovich, “On a problem of Monge,” *Dokl. Akad. Nauk*, Vol. 3, 225–226, 1948.

[13] Y. Linde and R. M. Gray, “A fake process approach to data compression,” *IEEE Trans. on Comm.*, Vol. COM–26, pp. 840–847, June 1978.

[14] M. Z. Mao, R.M. Gray and T. Linder, “Rate-Constrained Simulation and Source Coding IID Sources,” submitted to *IEEE Trans. Inform. Theory* for possible publication. Available at `http://ee.stanford.edu/~gray/trellis.pdf`.

[15] M. Marcellin and T. Fischer, “Trellis Coded Quantization of Memoryless and Gauss-Markov Sources,” *IEEE Trans. on Comm.*, Vol. 38, No. 1, pp. 92–93, January 1990.

[16] D. Ornstein, “An application of ergodic theory to probability theory,” *Annals of Probability*, 1:43–58, 1973.

[17] W. A. Pearlman, “Sliding-Block and Random Source Coding with Constrained Size Reproduction Alphabets,” *IEEE Transactions on Communications*, Vol. COM-30, NO. 8, Aug 1982, pp. 1859-1867.

[18] K. Rose, “A mapping approach to rate-distortion computation and analysis,” *IEEE Transactions on Information Theory*, Vol. 40, No. 6, pp.1939– 1952, November 1994.

[19] L. Stewart, R. M. Gray, and Y. Linde, “The design of trellis waveform coders,” *IEEE Transactions on Communications,* Vol. COM-30, No. 4, April 1982, pp. 702-710.

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

[21] C. Villani, *Optimal Transport, Old and New*, Springer, 2009.

[22] 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.

[23] S. G. Wilson and D. W. Lytle, “Trellis coding of continuous-amplitude memoryless sources,” *IEEE Trans. on Inform. Theory*, Vol. IT-23, pp. 404–409, May 1977.