Comparisons of Theoretical Limits for Source Coding With Practical Compression Algorithms

F. Pollara and S. Dolinar
Communications Systems Research Section

In this article, the performance achieved by some specific data compression algorithms is compared with absolute limits prescribed by rate distortion theory for Gaussian sources under the mean square error distortion criterion. These results show the gains available from source coding and can be used as a reference for the evaluation of future compression schemes. Some current schemes perform well, but there is still room for improvement.

I. Introduction

The theoretical limits on the performance of source and channel coding are well known for several source and channel models [1,2,5]. In this article, the authors calculate the theoretical limits for one- and two-dimensional Gauss-Markov sources used as models for planetary images. The formulas underlying these calculations are well known; the aims in this article are first to collect and graphically display these results, and then to compare them with the performance of specific data compression algorithms.

These results show the gains available from source coding and can be used as a reference for the evaluation of present and future compression schemes. These results also suggest that large improvements in information transmission in future missions can be achieved by advanced source coding.

II. Theoretical Rate Distortion Limits

The authors consider time-discrete continuous-amplitude sources that produce identically distributed output samples x governed by a probability distribution $P(x)$ with density $p(x)$. Each source sample x is reconstructed after source coding and decoding into a reconstructed sample y. The accuracy of reproduction is measured by a nonnegative function $d(x,y) = (x-y)^2$ called a squared error distortion measure. The average distortion D on a sequence of N samples is $(1/N)\sum_{i=0}^{N-1}(x_i - y_i)^2$ and is called mean square error (MSE) distortion.

A. One-Dimensional Gaussian Sources

For a Gaussian memoryless source, $p(x)$ is the Gaussian probability density with variance σ_x^2, and the rate distortion function for MSE distortion is [1]

$$R(D) = \frac{1}{2}\log_2 \frac{\sigma_x^2}{D}, \quad 0 \le D \le \sigma_x^2 \tag{1}$$

where the rate R is measured in bits/sample.

A time-discrete stationary Gaussian source with spectral density function

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \phi(n) e^{-jn\omega} \tag{2}$$

where $\phi(n)$ is the autocorrelation function, has a rate distortion $R(D)$ given in the parametric form [1]

$$D(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min[\theta, \Phi(\omega)] d\omega \tag{3}$$

and

$$R(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left[0, \frac{1}{2} \log_2 \frac{\Phi(\omega)}{\theta}\right] d\omega \tag{4}$$

where θ is the parameter.

Consider the special case of a first-order Gauss-Markov source of variance σ_x^2 with samples

$$x_i = \rho x_{i-1} + w_i \tag{5}$$

where $\{w_i\}$ is an independent, identically distributed (i.i.d.) zero-mean Gaussian sequence with variance $\sigma_w^2 = \sigma_x^2(1-\rho^2)$. This source will be called the one-dimensional causal model, or 1DC model, and is characterized by an exponentially decaying memory given by the autocorrelation function

$$\phi(n) = \sigma_x^2 |\rho|^n, \quad 0 \le \rho < 1 \tag{6}$$

which gives

$$\Phi(\omega) = \frac{\sigma_x^2(1-\rho^2)}{1 - 2\rho\cos\omega + \rho^2} \tag{7}$$

Incidentally, the power spectral density function is always easy to find, given the definition of the model that generates the samples $\{x_i\}$, as described in [3].

B. Two-Dimensional Gaussian Sources

The rate distortion function $R(D)$ for a two-dimensional Gaussian source is given by [7]

$$D(\theta) = (\frac{1}{2\pi})^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \min[\theta, \Phi(\omega_1, \omega_2)] d\omega_1 d\omega_2 \tag{8}$$

and

$$R(\theta) = (\frac{1}{2\pi})^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \max\left[0, \frac{1}{2} \log_2 \frac{\Phi(\omega_1, \omega_2)}{\theta}\right] d\omega_1 \omega_2 \tag{9}$$

A two-dimensional Gauss-Markov (autoregressive) causal source is defined by

$$x_{i,j} = \rho_1 x_{i-1,j} + \rho_2 x_{i,j-1} + \rho_{1,2} x_{i-1,j-1} + w_{i,j} \tag{10}$$

where $\{w_{i,j}\}$ is a two-dimensional i.i.d. zero-mean Gaussian sequence with variance σ_w^2. If $\rho_{1,2} = -\rho_1\rho_2$ is chosen, the source model in Eq. (10) becomes separable and will be called the two-dimensional causal (2DC) model. Then the variances of the sequences $\{w_{i,j}\}$ and $\{x_{i,j}\}$ are related by $\sigma_w^2 = \sigma_x^2(1-\rho_1^2)(1-\rho_2^2)$, and

$$\Phi(\omega_1, \omega_2) = \frac{\sigma_x^2(1-\rho_1^2)(1-\rho_2^2)}{(1 - 2\rho_1\cos\omega_1 + \rho_1^2)(1 - 2\rho_2\cos\omega_2 + \rho_2^2)} \tag{11}$$

This causal separable model has an autocorrelation function $\phi(n_1, n_2)$ given by

$$\phi(n_1, n_2) = \sigma_x^2 |\rho_1|^{n_1} |\rho_2|^{n_2} \tag{12}$$

which displays an undesirable nonisotropic behavior, as discussed later in this section.

Figure 1 shows the rate distortion functions for the 1DC model and the 2DC model with $\rho_1 = \rho_2 = \rho$ for several values of ρ. The values of the correlation coefficient ρ have been chosen to illustrate the effect of correlation on the rate necessary to represent the source. At low distortion, these values give rate distortion curves spaced by an integer number of bits from the curve for the memoryless source. Each successive correlation value in Fig. 1 represents (asymptotically for low distortion) one extra bit of information that can be extracted from each sample's correlation with its neighbors, and thus need not be spent to represent the source.

A more realistic model for images, the two-dimensional noncausal (2DNC) model, is given by

$$x_{i,j} = \alpha(x_{i,j-1} + x_{i,j+1} + x_{i-1,j} + x_{i+1,j}) + w_{i,j}, \quad |\alpha| < 1/4 \tag{13}$$

This is a noncausal model with a power spectral density

$$\Phi(\omega_1, \omega_2) = \frac{\sigma_w^2}{[1 - 2\alpha(\cos\omega_1 + \cos\omega_2)]^2} \tag{14}$$

where $\sigma_w^2 = \sigma_x^2 \eta$, and

$$\eta^{-1} = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{[1 - 2\alpha(\cos\omega_1 + \cos\omega_2)]^2} d\omega_1 d\omega_2 \tag{15}$$

The autocorrelation of this model was computed numerically, and it was found to behave almost isotropically at small displacements. Figure 2 shows a comparison of the autocorrelations for the causal and noncausal models for correlation coefficients ρ that are representative of typical planetary images. The function

$$\phi(n_1, n_2) = \sigma_x^2 |\rho|^{\sqrt{n_1^2 + n_2^2}} \tag{16}$$

is an example of exactly isotropic autocorrelation [4], but the authors do not presently know a model that realizes such an autocorrelation.

The qualitative behavior of the autocorrelation functions for the 2DC and 2DNC models is illustrated in the contour plots of Fig. 3. Note that for small values of n_1 and n_2 the contours for the 2DNC model are nearly circular, indicating that this model is nearly isotropic for small displacements.

The rate distortion functions for the 2DC and 2DNC models are shown in Fig. 4 with the same parameter values used in Fig. 2. Since the autocorrelation function for the 2DNC model decays more rapidly than for the 2DC model when both models have the same value of $\phi(1, 0)$ fixed, the rate distortion function of the 2DNC model lies above that of the 2DC model.

III. Quantization

Given a time-discrete continuous amplitude source, the simplest form of data compression is scalar (sample-by-sample) quantization. An M-level quantizer is a device with an input that can assume any real value x and an output y that can assume only M values $\{L_1, \cdots, L_M\}$. Usually, the number of levels is a power of 2, so that a B-bit quantizer has 2^B levels. Given the quantization thresholds $\{T_1, \cdots, T_{M-1}\}$, the output is $y = L_k$ if and only if $T_{k-1} < x \leq T_k$, $k = 1, \cdots, M$, where $T_0 = -\infty$ and $T_M = +\infty$. The input–output characteristic of a four-level quantizer is shown in Fig. 5.

Let $\{x_1, \cdots, x_N\}$ be a sequence of random samples generated by a source and let $\{y_1, \cdots, y_N\}$ be the corresponding quantized samples produced by an M-level quantizer. Then the quantized sequence has rate $B = \log_2 M$ bits and MSE distortion

$$D \triangleq \frac{1}{N}\sum_{i=1}^{N} E[(x_i - y_i)^2] = E[(x_i - y_i)^2]$$
$$= \sum_{k=1}^{M} \int_{T_{k-1}}^{T_k} (x - L_k)^2 p(x) dx \tag{17}$$

where $p(x)$ is the probability density of the source. Therefore, the M-level quantizer realizes the point (B, D) on the rate distortion plane. The optimum quantizer, which achieves the lowest possible MSE for given source statistics, has been determined in terms of the reproduction levels $\{L_k\}$ and the thresholds $\{T_k\}$ using an optimization technique developed by Lloyd and Max in 1960. If the quantizer is restricted to have equally spaced thresholds, i.e., a uniform quantizer with constant step size $T_k - T_{k-1}$ is considered, a slightly higher distortion for corresponding rates is obtained, as shown in Fig. 6 for the Gaussian memoryless source. An optimum uniform quantizer is a uniform quantizer that minimizes the MSE distortion.

Improved rate performance can be obtained by using entropy coding after quantization, since the probability $P_k = \Pr(y = L_k) = \int_{T_{k-1}}^{T_k} p(x) dx$ that a quantizer output will be L_k is not a constant (except for degenerate cases), and therefore the entropy of the quantized samples y is strictly less than B

$$H(y) = -\sum_{k=1}^{N} P_k \log_2 P_k < B \tag{18}$$

The entropy coded performance of the two quantizers considered above is also shown in Fig. 6, where it is apparent that the advantage of the Lloyd-Max quantizer over the uniform quantizer disappears after entropy coding. Results on entropy coded quantizers were obtained from the literature [3] and reproduced by computer simulation.

Instead of quantizing individual source samples, one could collect a whole vector $\mathbf{x} = (x_1, \cdots, x_n)$ and then

vector quantization. The performance of vector quantization methods will be discussed in a future article.

If one replaces the memoryless Gaussian source with a one-dimensional Gauss-Markov source with correlation coefficient ρ between successive samples (1DC model), a simple method to exploit the source memory is to take differences between successive quantized samples and then apply entropy coding. The performance of such a one-step predictor on samples produced by an optimum uniform quantizer is shown in Fig. 7.

In practice, the continuous amplitude source is initially quantized to B bits, typically 8 bits. In the following discussion of practical compression algorithms for images, it is assumed that the source has been quantized to 8 bits per sample by an optimum uniform quantizer.

IV. Comparisons of Practical Compression Algorithms and Theoretical Limits

The performance of specific compression algorithms designed for 8-bit input data can be measured experimentally by generating in software a Gauss-Markov random field according to one of the models described in Section II and by quantizing the resulting samples to 8 bits with an optimum uniform quantizer.

The entropy coded one-step predictor described in the previous section is a simple example of a practical compression scheme, and it is essentially the image compression scheme used in Voyager, where the source was initially quantized to 8 bits by the camera. The point denoted by 8 in the rate distortion plot of Fig. 7 represents the so-called lossless performance of such a scheme. This scheme performs reasonably well at low distortions (as compared with the rate distortion function) when it is applied to the one-dimensional source 1DC. One will see that its performance is no longer attractive when applied to two-dimensional sources 2DC or 2DNC.

The proposed Joint Photographic Expert Group (JPEG) image compression standard [6] uses, in its baseline version, discrete cosine transform (DCT) processing, quantization, and Huffman coding. The performance of this compression scheme on the 2DC model has been evaluated and compared with the rate distortion limits in Fig. 8. The performance of the entropy coded one-step predictor on the 2DC model is also shown in Fig. 8 for comparison. For most science purposes, a typical planetary image is considered acceptable at normalized distortions D/σ_x^2 up to approximately 10^{-2}, corresponding to about 5 gray levels of rms error out of 256 levels for typical images. In this range of interest, the JPEG scheme is superior to the entropy coded predictor, but the theoretical limit leaves ample space for improvements. The performances of the JPEG scheme and the entropy coded one-step predictor on the 2DNC model are compared in Fig. 9.

V. Conclusion

The theoretical limits computed in this article and the experimental results on source models verify the gains available by source coding, and can be used as a reference for the evaluation of present and future compression schemes. These results also suggest that large improvements in information transmission in future missions can be achieved by advanced source coding.

Mathematical source models studied in this article include both relatively simple one- and two-dimensional causal Gauss-Markov models and a two-dimensional noncausal model whose nearly isotropic correlation function more closely resembles that of real images.

More work is necessary in relating the mathematical models to actual image sources, in evaluating the performance of other practical compression schemes, and in understanding the actual quantization performed in the camera.

References

[1] T. Berger, *Rate Distortion Theory*, Englewood Cliffs, New Jersey: Prentice Hall, 1971.

[2] S. Dolinar and F. Pollara, "The Theoretical Limits of Source and Channel Coding," *TDA Progress Report 42-102*, vol. April–June 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 62–72, August 15, 1990.

[3] A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, New Jersey: Prentice Hall, 1989.

[4] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, New Jersey: Prentice Hall, 1984.

[5] R. J. McEliece, *The Theory of Information and Coding*, Reading, Massachusetts: Addison Wesley, 1977.

[6] F. Pollara and S. Arnold, "Emerging Standards for Still Image Compression: A Software Implementation and Simulation Study," *TDA Progress Report 42-104*, vol. October–December 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 98–102, February 15, 1991.

[7] J. A. Stuller and B. Kurz, "Intraframe Sequential Picture Coding," *IEEE Transactions on Communications*, vol. COM-25, no. 5, pp. 485–495, May 1977.

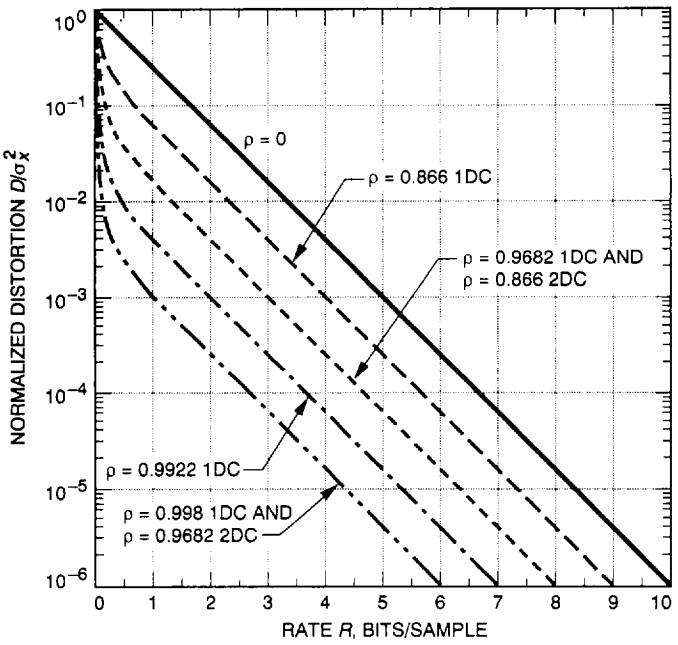

Fig. 1. Rate distortion functions for 1DC and 2DC models.

(a)

$\rho = \rho_1 = \rho_2 = 0.9$

n_2 \ n_1	0	1	2
0	1.000	0.900	0.810
1	0.900	0.810	0.729
2	0.810	0.729	0.656

(b)

$\alpha = 0.245$

n_2 \ n_1	0	1	2
0	1.000	0.904	0.772
1	0.904	0.845	0.737
2	0.772	0.737	0.660

$\rho = \rho_1 = \rho_2 = 0.98$

n_2 \ n_1	0	1	2
0	1.000	0.980	0.960
1	0.980	0.960	0.941
2	0.960	0.941	0.922

$\alpha = 0.249$

n_2 \ n_1	0	1	2
0	1.000	0.974	0.929
1	0.974	0.956	0.916
2	0.929	0.916	0.884

Fig. 2. Two-dimensional normalized autocorrelation functions $\phi(n_1, n_2)/\sigma_x^2$: (a) 2DC model and (b) 2DNC model.

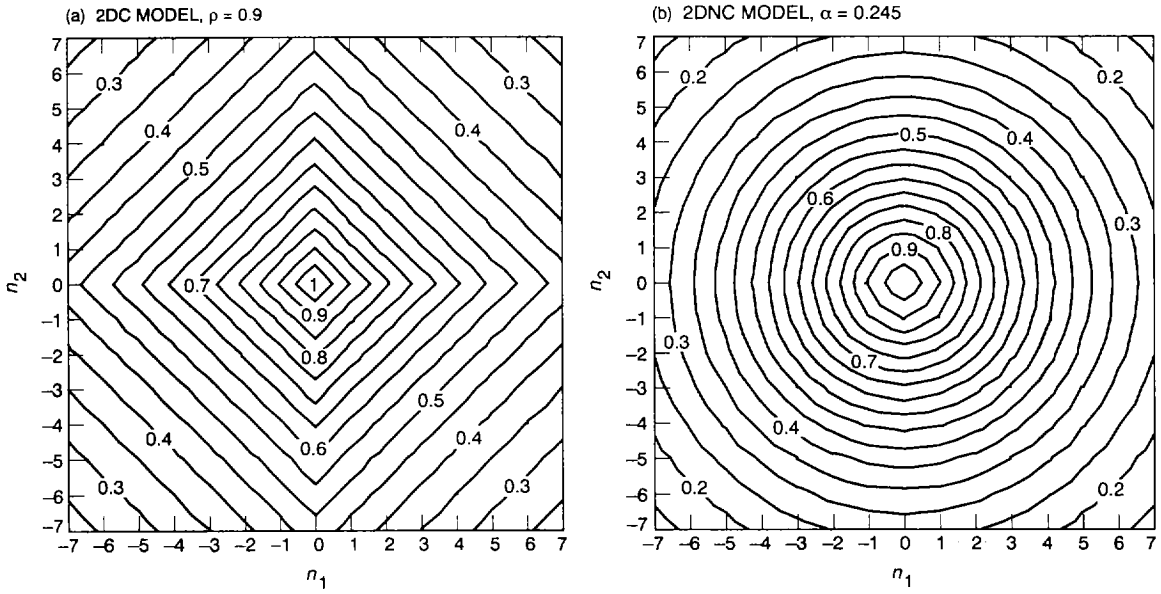

Fig. 3. Contour plots: (a) causal and (b) noncausal autocorrelations.

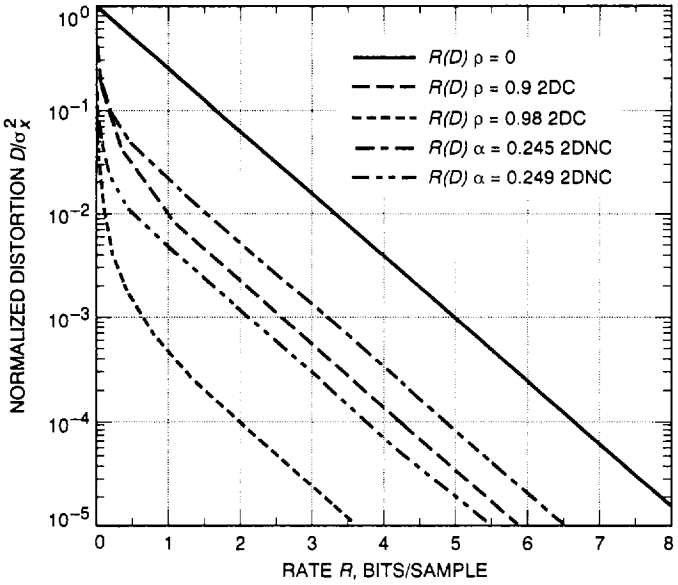

Fig. 4. Rate distortion functions for causal and noncausal two-dimensional models.

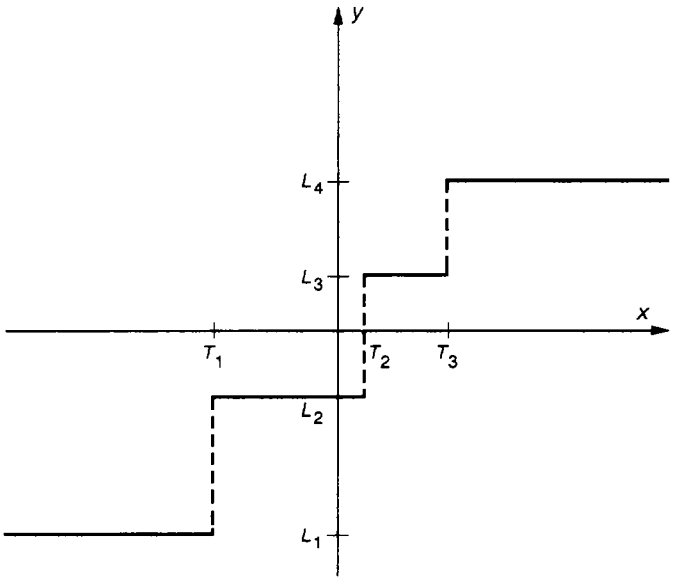

Fig. 5. Input–output characteristic of a four-level quantizer.

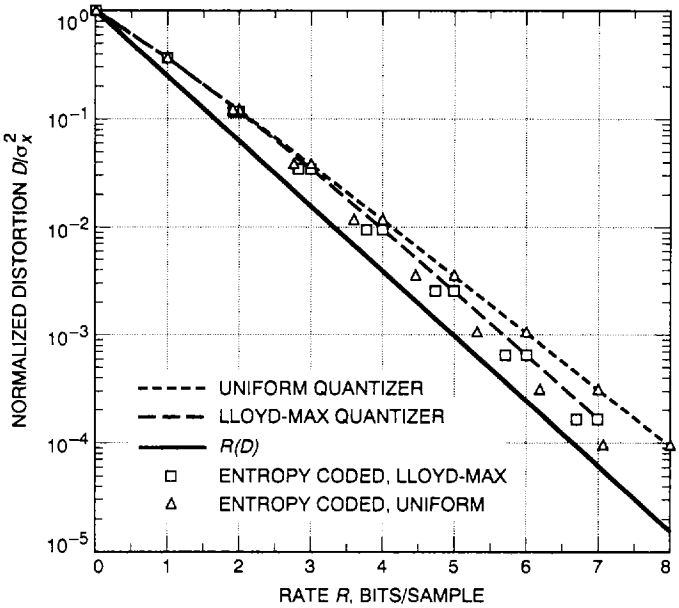

Fig. 6. Performance of quantization schemes for memoryless Gaussian sources.

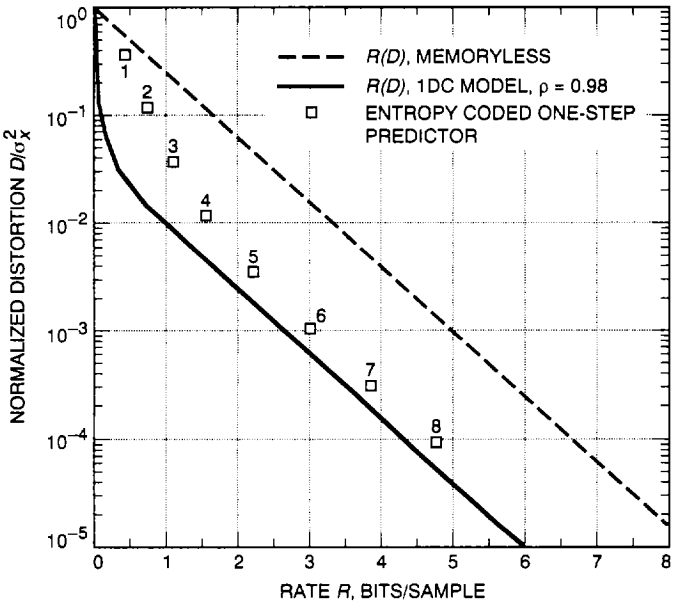

Fig. 7. Performance of entropy coded one-step predictor on Gauss-Markov source with $\rho = 0.98$.

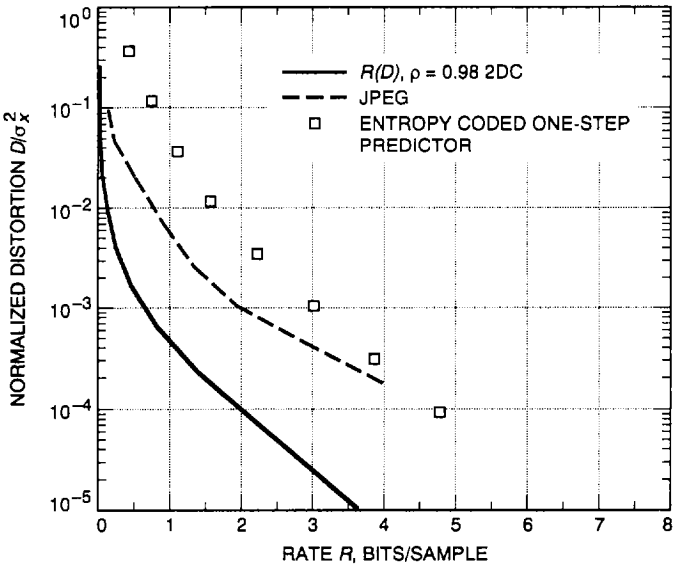

Fig. 8. Comparisons of practical compression algorithms and theoretical limits (2DC model).

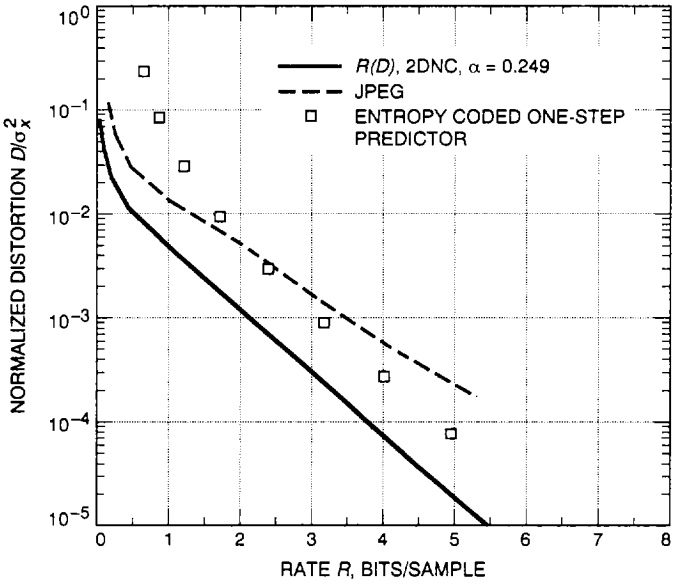

Fig. 9. Comparisons of practical compression algorithms and theoretical limits (2DNC model).