Rate Distortion Behavior of Sparse Sources

Claudio Weidmann, Martin Vetterli

October 12, 2001

Abstract

This paper studies the rate distortion behavior of sparse memoryless sources that serve as models of sparse signal representations. For the Hamming distortion criterion, $R(D)$ is shown to be essentially linear. For the mean squared error measure, two models are analyzed: the mixed discrete/continuous spike processes and Gaussian mixtures. The latter are shown to be a better model for "natural" data such as sparse wavelet coefficients. Finally, the geometric mean of a continuous random variable is introduced as a sparseness measure. It yields upper and lower bounds on the entropy and thus characterizes high-rate $R(D)$.

Keywords

Sparse signal representations, rate distortion theory, memoryless systems, entropy, transform coding.

I. Introduction

THE success of wavelet based coding, especially in image compression, is often attributed to the ability of wavelets to "isolate" singularities, something Fourier bases fail to do efficiently. Thus, a piecewise smooth signal is mapped through the wavelet transform into a sparse set of non-zero transform coefficients, namely coefficients around discontinuities as well as coefficients representing the general trend of the signal. While this behavior is well understood in terms of nonlinear approximation power (approximation by N largest terms of the wavelet transform, see [1] for a thorough treatment), the rate-distortion behavior is more open.

The work by Mallat and Falzon [2] was the first to analyze the low-rate behavior of transform image coding, and showed the very different behavior with respect to classic, Karhunen-Loève

The material in this paper was presented in part at the Data Compression Conference, Snowbird UT, March 1999 and 2000, and at the IEEE International Symposium on Information Theory, Washington DC, June 2001. Part of this work stems from the Ph.D. thesis of the first author and was supported by an ETHZ/EPFL fellowship.

The authors are with the Audiovisual Communications Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. E-mail: Claudio.Weidmann, Martin.Vetterli@epfl.ch.

M. Vetterli is also with the Department of EECS, UC Berkeley, Berkeley CA 94720.

transform (KLT) theory. In essence, at low rates, only few wavelet coefficients are involved in the approximation of piecewise smooth functions, leading to a steeper decline of the rate-distortion function as compared to the classic exponential decay in the case of Gauss-Markov processes and the KLT. This result had been observed experimentally in low-rate image coding (see Figure 1 for an example).

The results above indicate the interest to understand more fully the rate-distortion behavior of sparse vectors. The wavelet transform being a unitary map, it is sufficient to get bounds on the rate-distortion function of sparse sources in order to understand the compression of sources that are "sparsified" by the wavelet transform, like piecewise smooth functions.

It is probably worthwhile to contrast the KLT on jointly Gaussian processes with the wavelet transform on piecewise smooth processes. In the KLT case, the optimal strategy is waterfilling [3], and the approximation process is linear (up to quantization). In the wavelet transform approach, the approximation is non-linear, and a key element of efficient compression is to "point to" the important coefficients (for example, many data structures have been proposed just for this, e.g. zero trees [4]). This points again to the importance of "location" in compressing vectors with few important coefficients.

In this paper, we consider various forms of sparse vectors, where both position and value are important. The first case, in Section II, deals with pure position coding by considering binary vectors and Hamming distortion. In the deterministic case, when the number of non-zero entries is known a priori, it is possible to give closed form rate-distortion functions. Interestingly, for sparse spikes, the R(D) function is "almost" linear. In the non-deterministic case of a Bernoulli-p source, it is shown in Theorem 1 that the normalized rate-distortion function is asymptotically linear as $p \to 0$.

In Section III, a mixed discrete/continuous spike process is considered. This Bernoulli-Gaussian spike process uses a Bernoulli process to turn a normal random variable on or off. So both position and value are important. The rate distortion behavior of the spike process is characterized using classification-based upper bounds. However, the results do not match experimental rate-distortion curves closely because of the mixed discrete/continuous nature.

This leads to consider Gaussian mixtures in Section IV, where a hidden process picks one of two normal random variables with different variances. Both upper and lower bounds show

the knee in the rate-distortion curve that is typical for such mixtures. A notion akin to the classic coding gain of transform coding is introduced, which is based on magnitude classifying quantization. Instead of separately considering the transform coefficients, they can be mixed without incurring much loss.

Finally, Section V considers the geometric mean of a source as a sparseness measure, which is used to bound the coding gain of Section IV. Additionally, the geometric mean yields lower and upper bounds on the source entropy. Therefore it is a good means to characterize the high-rate rate distortion behavior of sparse sources.

II. Spike Position Encoding

Consider a source emitting sparse random vectors of length N in which most components are zero, except for a few *spikes* that stick out. In this section, we are only interested in the positions of the nonzero values (the spikes), therefore we can restrict ourselves to binary vectors. A lossy encoder maps a source vector $\boldsymbol{X}$ to a reconstructed version $\hat{\boldsymbol{X}}$. The fidelity of this approximation is measured by the Hamming distance:

$$d_H(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \sum_{i=1}^{N} [1 - \delta(X_i - \hat{X}_i)]. \tag{1}$$

This is equivalent to a frequency of error criterion where both types of errors have the same cost (coding a spike when there is none and vice-versa). The rate distortion function $R(D)$ gives the minimum rate R necessary to encode the source with fidelity D. In the following, we will first consider a purely combinatorial setting, where exactly K out of N positions are equal to one, with a uniform prior on the $\binom{N}{K}$ possible combinations. Hence the problem looses its dimensionality and can actually be solved with the methods for discrete memoryless sources which are summarized in the appendix.

A. Single Spike

The source $\boldsymbol{X}$ is equivalent to a i.i.d. uniform source U with alphabet $\mathcal{U} = \{1, 2, \dots, N\}$. Using the standard basis vectors $\mathbf{e}_i$ we can write $\boldsymbol{X} = \mathbf{e}_U$. It can be shown (see Theorem 10 from [5] in the appendix) that just one additional reconstruction letter is needed to achieve the rate distortion bound, and it will map to the all-zero vector $\mathbf{0}$. To see that it can only be the all-zero vector, consider the source alphabet $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N\}$, which consists of all vectors of

Hamming weight one. Any other non-zero vector will be at Hamming distance one or more from these vectors and thus can only worsen the distortion achieved by the all-zero vector, i.e. exactly one. If we define $\widehat{\mathcal{U}} = \mathcal{U} \cup \{0\}$ and $\mathbf{e}_0 = \mathbf{0}$, then everything fits nicely. Using $\hat{u} = 0$ corresponds to not coding the position. We get the following distortion measure:

$$\rho(u; \hat{u}) = d_H(\mathbf{e}_u, \mathbf{e}_{\hat{u}}) = 2[1 - \delta(u - \hat{u})] - \delta(\hat{u}) \tag{2}$$

Thus "giving the right answer" has zero distortion, a wrong answer two, and not answering costs one distortion unit.

Proposition 1 *The rate distortion function for a single spike in $N \geq 2$ equiprobable positions with the Hamming distortion criterion (1) is*

$$R(D) = \begin{cases} (1 - D)\log(N - 1) & \text{if } \frac{2}{N} < D \leq 1, \\ \log N - \frac{D}{2}\log(N - 1) - h_b\left(\frac{D}{2}\right) & \text{if } 0 \leq D \leq \frac{2}{N}. \end{cases} \tag{3}$$

Proof: The following derivation relies heavily on the rate distortion results for discrete memoryless sources summarized in the appendix. There it is shown that $R(D)$ can be computed by solving a set of equations involving the marginal (random codebook) distribution $Q(k)$ on the reconstruction alphabet. The symmetry of the input distribution, $P(j) = 1/N$ $(j = 1, \ldots, N)$, suggests the following marginal distribution (with a slight abuse of notation):

$$Q = (q_0, q_1 = q_2 = \ldots = q_N = \frac{1 - q_0}{N}). \tag{4}$$

Let us first assume that $q_k > 0$ holds for all k. Then the $N+1$ conditions (61) from the Appendix have to be met. We make the substitution $\beta = e^{-\lambda}$ and insert our $Q(k)$ into the equation, first for $k \neq 0$:

$$\begin{aligned} \frac{\beta^0}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} + \frac{(N-1)\beta^2}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} &= \frac{1}{P(j)} = N \\ &\vdots \\ q_0((N-1)\beta^2 - N\beta + 1) &= 0. \end{aligned} \tag{5}$$

For $k = 0$ we get almost the same equation:

$$\frac{N\beta^1}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} = \frac{1}{P(j)} = N$$

$$\vdots$$

$$(1 - q_0)((N-1)\beta^2 - N\beta + 1) = 0. \tag{6}$$

The solution $\beta = 1$ corresponds to the point $(0, D_{max})$ (with $D_{max} = 1$) in the (R, D) plane, which is achieved by setting $q_0 = 1$. Therefore the interesting solution is $\beta = 1/(N-1)$, which when inserted into (60) yields

$$Q(k|j) = q_k(N-1)^{1-\rho(j,k)}. \tag{7}$$

Putting (7) into (58) we get the average distortion $d(Q) = 1 - \frac{N-2}{N}(1 - q_0)$ and from (59) the rate $I(Q) = \frac{N-2}{N}(1 - q_0)\log(N-1)$. Noting that these hold for $q_0 > 0$, we combine them to eliminate q_0 and get

$$D(R) = 1 - \frac{R}{\log(N-1)} \qquad \text{for } R < \frac{N-2}{N}\log(N-1). \tag{8}$$

This proves the first part of equation (3). When R reaches its upper bound in (8), D reaches $2/N$ and we have $q_0 = 0$. At that point, equation (5) will be satisfied for all β. According to condition (62), equation (6) now becomes an inequality:

$$(N-1)\beta^2 - N\beta + 1 \geq 0. \tag{9}$$

This is satisfied by $\beta \geq 1$ or $\beta \leq \frac{1}{N-1}$, which is equivalent to $\lambda \geq \log(N-1)$. The first solution ($\beta \geq 1$) can be discarded, since it would result in $D(R)$ being larger than 1 and discontinuous. The conditional distribution parameterized by β is

$$Q(k|j) = \begin{cases} 0, & k = 0 \\ \frac{\beta^{\rho(j,k)}}{1+(N-1)\beta^2}, & k \neq 0 \end{cases} \tag{10}$$

As before, we put this into (58) to get $d(Q) = \frac{2(N-1)\beta^2}{1+(N-1)\beta^2}$ and into (59) yielding

$$I(Q) = \log N - \frac{(N-1)\beta^2}{1+(N-1)\beta^2}\log(N-1) - h_b\left(\frac{1}{1+(N-1)\beta^2}\right),$$

where $h_b(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. Eliminating β from the last two equations yields the second part of equation (3). ■

Figure 2 shows a set of typical $R(D)$ functions. As N grows large, the linear segment dominates the rate distortion characteristics. Further we observe that in the special case $N = 2$ the solution degrades to the $R(D)$ function of a binary symmetric source (with doubled distortion).

B. Multiple Spikes

Now we consider a source emitting one of the $\binom{N}{K}$ binary vectors of length N and Hamming weight K. We will again assume that all source letters are equally probable, and that N and K are given. We look only at the case where the number of 1's is $K \leq N/2$, since the other case ($N/2 \leq K \leq N$) is complementary.

The analysis is simplified by the fact that the set of source vectors of weight K forms a group code under permutation. Under the action of the symmetric group S_N, any vector of the set will again yield the whole set. The code is thus geometrically uniform, i.e. the distance (distortion) profile looks the same from any vector in the set. By decomposing permutations into transpositions, one establishes that the distances will always be integer multiples of two. Assuming that $K \leq N/2$, there are exactly

$$w_d = \binom{K}{d}\binom{N-K}{d}, \quad d = 0, \ldots, K \tag{11}$$

vectors at Hamming distance $2d$ from a given vector. The following identity will also be very helpful in our development:

$$\sum_{d=0}^{K} w_d = \sum_{d=0}^{K} \binom{K}{d}\binom{N-K}{d} = \binom{N}{K}. \tag{12}$$

As in the single spike case, the reconstruction alphabet consists of the source alphabet plus the zero vector, to which we assign the probability q_0 as before. To compute the slope of the linear part of the rate distortion curve we have to solve (compare with (5, 6))

$$\sum_{d=0}^{K} w_d \beta^{2d} - \binom{N}{K}\beta^K = 0. \tag{13}$$

The solution $\beta = 1$ corresponds again to the maximum distortion, $D = K$. We will now assume that somehow we found the interesting root β_0 with $0 < \beta_0 < 1$ (for $K = 2$ it is $\beta_0 = \binom{N-2}{2}^{-1/2}$, for larger K it can be computed numerically). Then the linear part of the rate distortion function will be

$$R(D) = (D - K)\log\beta_0, \quad D(\beta_0) < D < K \tag{14}$$

where the bounds on D guarantee $0 < q_0 < 1$ ($D(\beta_0)$ is defined below in (16a)). For $q_0 = 0$, any $\beta \leq \beta_0$ will satisfy the Kuhn-Tucker conditions. We define a pseudo-distribution

$$b_d = \frac{w_d \beta^{2d}}{\sum_{d'=0}^{K} w_{d'} \beta^{2d'}}, \quad d = 0, \ldots, K. \tag{15}$$

After some calculations, we get a parametric expression for the rate-distortion curve for $0 < \beta < \beta_0$:

$$D(\beta) = \sum_{d=1}^{K} b_d 2d, \tag{16a}$$

$$R(\beta) = \log \binom{N}{K} + \sum_{d=0}^{K} b_d \log b_d - \sum_{d=0}^{K} b_d \log w_d. \tag{16b}$$

The middle term in the expression for R is the negative entropy of the pseudo-distribution b_d (compare with (3)). Figure 3 shows that for sparse spikes (small K/N) the linear segment again dominates the rate distortion behavior. The consequence of this "almost linear" $R(D)$ behavior for sparse spikes is the following: to build a close to optimal encoder for intermediate rates $0 < R < \log \binom{N}{K}$, we can simply multiplex between a rate 0 code (no spikes coded) and one with rate $\log \binom{N}{K}$ (all K spikes coded exactly). Put otherwise, if we have a bit budget to be spent in coding a sparse binary vector, we can simply go ahead and code the exact positions of the ones (the spikes) until we run out of bits.

These results can also be used to derive the asymptotic operational rate distortion function of a simple two-pass universal lossy source coder: first, the number of ones (K) in a block of length N is determined and sent to the decoder using at most $\log_2 N$ bits. Then a code for a weight K vector is used. For $N \to \infty$, we approach the above rate distortion functions with a redundancy of $\frac{\log_2 N}{N}$ bits per sample. (In view of the results for universal lossless source coding [6], we expect that this redundancy could be halved.)

C. Nondeterministic case

The above results should be compared with the rate distortion function of a binary memoryless source (BMS) with $p = \Pr\{X = 1\} = K/N$, corresponding to the nondeterministic situation where only the average number of spikes is known a priori.

Theorem 1 *Consider a Bernoulli-p source ($p \leq \frac{1}{2}$ w.l.o.g.) with normalized Hamming distortion $d = D/p$. Then the normalized rate distortion function is asymptotically linear when*

$p \to 0$:

$$\lim_{p\to 0} \frac{R(d)}{h_b(p)} = 1 - d, \quad 0 \le d \le 1 \tag{17}$$

Proof: The rate distortion function for a BMS is $R(D) = h_b(p) - h_b(D)$ for $D \le p \le \frac{1}{2}$. Therefore

$$\begin{aligned} \frac{R(d)}{h_b(p)} &= 1 - \frac{h_b(pd)}{h_b(p)} \\ &= 1 - \frac{pd\log(pd)+(1-pd)\log(1-pd)}{p\log(p)+(1-p)\log(1-p)}, \end{aligned}$$

from which

$$\begin{aligned} \lim_{p\to 0} \frac{R(d)}{h_b(p)} &= 1 - \lim_{p\to 0} \frac{d\log(pd)-d\log(1-pd)}{\log p - \log(1-p)} \\ &= 1 - \lim_{p\to 0} \frac{d/p + d^2/(1-pd)}{1/p + 1/(1-pd)} \\ &= 1 - d \end{aligned}$$

■

Theorem 1 shows that if we normalize the rate and the distortion by their maxima, $h_b(p)$ and p, respectively, the rate distortion function becomes linear for sparse sources ($p \to 0$).

III. Scalar-valued Spike Processes

The previous section considered only the spike positions using the Hamming distortion measure. Now we also assign a scalar *value* to each spike, and the distortion will be measured by the mean squared error. Moreover, we abandon the setting "K spikes in N positions" in favor of a less deterministic model: a (scalar-valued) spike process is simply the product of a binary-$\{0, 1\}$ source with a memoryless real-valued source. Here we consider only Gaussian values, because they serve as the usual worst-case benchmark. The binary source simply "switches the value source on or off".

Definition 1 (Bernoulli-Gaussian (BG) spike process) *An i.i.d. Bernoulli-Gaussian spike source emits a memoryless random variable X that is the product of a binary random variable with $\Pr\{X = 1\} = p$ and $\Pr\{X = 0\} = 1-p$ and a zero mean Gaussian with variance σ^2. Using the $\delta(\cdot)$ distribution, the pdf of the BG spike can be written as*

$$f(x) = (1-p)\delta(x) + p\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}. \tag{18}$$

This pdf can also be seen as a mixture of two zero mean Gaussian random variables, with one of them having zero variance (a special case of the model that will be studied in the next section). In order to characterize the rate distortion behavior of the BG spike we use the upper bounds presented in [7], [8], which are obtained by classifying the magnitudes of the source samples using a threshold t and applying the Gaussian upper bound to each of the two classes.

These bounds are upper bounds on the operational rate distortion function $\delta(R)$ of *magnitude classifying quantization* (MCQ), which is a quantization method where the classification side information is used to switch between two codebooks. A low-rate bound is obtained by upper bounding only the samples with magnitude above threshold (also called *significant* samples), while the other samples are quantized to zero, thus yielding a distortion floor.

Theorem 2 [8] (Low-Rate Bound) *The distortion rate function of a memoryless source with symmetric pdf $f(x)$ and variance σ^2 is upper bounded by*

$$D(R) \leq B(t,R) = A(t)\left[\exp\left(-2\tfrac{R-h_b(\mu(t))}{\mu(t)}\right)-1\right]+\sigma^2, \quad \forall\, t \geq 0, \tag{19}$$

where the incomplete moments $\mu(t) = \Pr\{|X| \geq t\} = 2\int_t^\infty f(x)\,\mathrm{d}x$ and $A(t) = \mu(t)\,\mathrm{E}[X^2\,||X| \geq t] = 2\int_t^\infty f(x)x^2\,\mathrm{d}x$ are the ratio of significant samples and their unnormalized variance, respectively (note that $A(0) = \sigma^2$). In the neighborhood of a fixed threshold t the tightest bound is

$$D(R^*(t)) \leq B(t, R^*(t)), \quad \forall\, t \geq 0 : \exists R^*(t) \tag{20}$$

with the rate $R^*(t)$ given by

$$R^*(t) = h_b(\mu(t)) - \tfrac{1}{2}\mu(t)\left[2h_b'(\mu(t)) + \gamma(t) + W_{-1}\left(-\gamma(t)e^{-2h_b'(\mu(t))-\gamma(t)}\right)\right], \tag{21}$$

where γ is the reciprocal normalized tail variance $\gamma(t) = \frac{\mu(t)}{A(t)}t^2 = \frac{t^2}{\mathrm{E}[\mathrm{X}^2|\,\mathrm{X}\geq\mathrm{t}]}$ and W_{-1} is the second real branch of Lambert's W function, taking values on $[-1,-\infty)$. ($W(x)$ solves $W(x)e^{W(x)} = x$.)

We can use (20) to trace an upper bound on $D(R)$ by sweeping the threshold $t = 0\ldots\infty$. If we also consider the insignificant samples (below threshold), a high-rate bound results.

Theorem 3 [8] (High-Rate Bound) *Let the variances of the insignificant and the significant samples be $\sigma_0^2(t) = \mathrm{E}[X^2\,||X| < t] = \frac{\sigma^2 - A(t)}{1-\mu(t)}$ and $\sigma_1^2(t) = \mathrm{E}[X^2\,||X| \geq t] = \frac{A(t)}{\mu(t)}$, respectively.*

Then for all $R \geq R_{min}(t) = h_b(\mu(t)) + \frac{1}{2}\log\frac{\sigma_1^2(t)}{\sigma_0^2(t)}$, *the distortion rate function of a memoryless source is upper bounded by*

$$D(R) \leq B_{hr}(t,R) = c(t)\sigma^2 e^{-2R}, \tag{22}$$

where

$$c(t) = \exp\left(2h_b(\mu(t)) + [1-\mu(t)]\log\frac{1-A(t)/\sigma^2}{1-\mu(t)} + \mu(t)\log\frac{A(t)/\sigma^2}{\mu(t)}\right). \tag{23}$$

The best asymptotic upper bound for $R \to \infty$ *is obtained by numerically searching the* $t_0 \in [0,\infty)$ *that minimizes* $c(t)$. *Since* $\lim_{t\to 0+} c(t) = 1$, *the Gaussian upper bound is always a member of this family.*

The low-rate and high-rate bounds coincide in the minimum of the latter, i.e. as expected there is a smooth transition between the two bounds. For proofs, see [8]. We remark that results by Sakrison [9] and Gish-Pierce [10] imply that the operational distortion rate function $\delta(R)$ of a magnitude classifier followed by a Gaussian scalar quantizer (adapted to the class variance) will be at most a factor of $\pi e/6$ (1.53 dB) above these bounds. Actually, this gap is even smaller at low rates, since the "minimum" distortion $D(R_0)|_{R_0=0} = \sigma_0^2$ is trivially achieved for the insignificant samples.

The low-rate bound can be easily evaluated for the BG spike if one replaces t by $t+\epsilon$ in the lower integration boundaries, with an arbitrarily small number $\epsilon > 0$. By doing so, we exclude the Dirac $(1-p)\delta(x)$ from the integral, and hence we have $\mu(t) \leq p$ for all $t \geq 0$. This is obviously correct, since we never have to code the value of a spike with zero amplitude.

Figure 4 shows the low-rate bound and the empirical $D(R)$ for $p = 0.11$. The asymptote shown is actually the trivial upper bound $B(0,R)$, i.e. when all spikes are coded (thus at least $R = h_b(0.11) = 0.5$ bits are required before the distortion starts decreasing). The figure illustrates the change in $D(R)$ behavior between low and high rates that is typical of spike processes, regardless whether the continuous part of the pdf is a Gaussian or some other density. In particular, the asymptotic distortion decay is of the order of $-6/p$ dB per bit, which can be much steeper than the -6 dB typical of random variables with an absolutely continuous distribution.

In fact, spikes are *mixed* random variables that have both a discrete and a continuous part and for which most results in "standard" rate distortion theory do not hold. Their entropy

cannot be computed with the usual integral, but only via mutual information conditioned on the discrete part [11, Ch. 2]. With this trick, Rosenthal and Binia were able to derive the asymptotic rate distortion behavior of mixed random variables [12]. Their results coincide with our asymptotic upper bound $B(0, R)$ if the continuous part is Gaussian, otherwise their result is obviously tighter. These results were later extended to the vector case by Györgi et al. [13].

A natural extension of the memoryless spike model is to consider bursts of spikes. To model such a bursty behavior we can simply replace the Bernoulli process by a first order binary Markov process, with states S_0 (no spike) and S_1 (spike). An example of this has been studied in [8].

We proposed the spike process as a model for sparse transform coefficients. However, comparing with Figure 1 we see that its $D(R)$ behavior is very different from the one observed in actual image coders. The reason lies in the mixed nature of the model: at large rates, we have to spend the rate to code the discrete part, but in turn the distortion decay will be steeper, inversely proportional to the probability of nonzero samples. Conversely, by the tightness of the Shannon lower bound, a continuous random variable cannot have an asymptotic distortion decay other than the well known -6 dB per bit. Thus the spike process is not suited to model the coefficients of a transformed continuous random process.

However, there are other applications, such as using the spike as a benchmark for sparsifying transforms. For example, the KLT of a spike process will be dense, showing that the KLT is not optimal in terms of sparseness [14]. The work by Saito et al. is a further exploration of this direction [15].

IV. Gaussian Mixture Model

As became clear in the above discussion, continuous densities are more appropriate for modeling sparse transform coefficients. One of the more common approaches to density estimation is based on Gaussian mixtures. In this section we will analyze a simple i.i.d. Gaussian mixture model, where a hidden binary memoryless source picks one of two zero mean memoryless Gaussian sources. This is a generalization of the spike model, where one source had zero variance. The model pdf is:

$$f(x) = pf(x|S = 1) + (1 - p)f(x|S = 2) \tag{24}$$

where S is the hidden state selecting a source, and

$$f(x|S=i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-x^2/2\sigma_i^2}. \tag{25}$$

Such models have been used quite successfully in various applications, see e.g. [16] and references therein. To get realistic estimates for the parameters, we used the EM algorithm on the wavelet coefficients of the Lena image (transformed with the classic 9/7 biorthogonal wavelet).

Plots of the bounds (20) and (22) appear in Figure 5 together with the empirical $D(R)$ computed with Blahut's algorithm. Up to the knee, which is typical for image coding $D(R)$, the distortion decays faster than -6 db/bit. This means mainly that the sparse coefficients from the high variance source are retained by the thresholding operation. At higher rates, the coefficients from the low variance source also start being significant. If the model (24) is extended to three or more Gaussian components, the knee in $D(R)$ becomes rounder, but the basic behavior is unchanged. From these observations we can reach two conclusions: first, two-component Gaussian mixtures suffice to capture the essential features of image coding $D(R)$, and second, the rate $R_{min}(t_0)$ in the high-rate bound (Theorem 3) marks the beginning of the high-rate compression regime.

Gaussian mixture models have often been used in image compression, for example a classification approach has been proposed in [17]. The authors consider the joint numerical optimization of the classifier and (high-rate) uniform quantizers for each of the N classes. Their simulation results indicate that for typical image data $N = 2$ classes yield a substantial improvement over a single class. Adding more classes gives only minor gains over $N = 2$, which supports our observation that a two-component Gaussian mixture is a good basic model for wavelet coefficients.

A. Oracle Lower Bound on $D(R)$ of Gaussian Mixtures

Since Gaussian mixtures are a popular tool to approximate unknown densities, it is useful to also have a lower bound on their rate distortion function. The Gaussian mixture source can be viewed as a discrete memoryless source S that switches between $|\mathcal{S}|$ Gaussian sources $\mathcal{N}(m_s, \sigma_s^2)$ with selection probabilities $w_s = Pr\{S = s\}$. A lower bound on $D(R)$ is found by assuming that an oracle provides the hidden state variable S to the source encoder. Since $S \to X \to \widehat{X}$

form a Markov chain, we have

$$I(X;\widehat{X}|S) \leq I(X;\widehat{X}). \tag{26}$$

We observe that $R_{lb}(D) = \min_{p(\hat{x}|x,s)\in Q_D} I(X;\widehat{X}|S)$ (with $Q_D = \{p(\hat{x}|x,s) : \mathrm{E}(\mathrm{X}-\widehat{\mathrm{X}})^2 \leq \mathrm{D}\}$) can be computed exactly by solving the following standard rate allocation problem:

$$D_{lb}(R_{lb}) = \min_{\{R_s\}} \sum w_s \sigma_s^2 2^{-2R_s} \tag{27}$$

subject to

$$\sum w_s R_s = R_{lb} \text{ and } R_s \geq 0. \tag{28}$$

This yields the lower bound $D(R) \geq D_{lb}(R)$, which can be seen as a special case of a conditional rate distortion function [18].

Figure 5 shows the lower bound (27), together with the upper bounds from the previous section and the (R, D) points achieved by a scalar bitplane quantizer (applied to $3 \cdot 10^5$ pseudo-random samples from the mixture source; significance maps are entropy coded, sign and refinements bits left uncoded). At low rates, thresholding with simple scalar quantization performs very close to the R/D optimum.

B. Coding Gain Revisited

In linear transform coding, the *coding gain* measures the compression gain of a transform coding system with quantizer bit allocation compared to a single scalar quantizer without transform. Here we show how the high-rate upper bound (Theorem 3) leads to an expression that is reminiscent of the coding gain of a two-dimensional transform coding system.

Let us quickly go through the derivation of the classical transform coding gain. Consider a real-valued, time-discrete, stationary and ergodic process $\{X_k\}$ with mean zero and variance σ^2. The samples are grouped into blocks $\boldsymbol{X} = [X_{i=1}^N]$ of length N and transformed with an orthonormal transform: $\boldsymbol{Y} = T\boldsymbol{X}$. By Parseval's equality, the quantization error in the signal domain will be equal to the error in the transform domain:

$$\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2.$$

Also, the average variance of the transform coefficients Y_i is equal to the variance of X:

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, Y_i^2 = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, X_i^2 = \sigma^2.$$

This holds (by linearity of expectation) assuming zero mean and can be easily extended to the non-zero mean case. Let $\sigma_i^2 = \mathrm{E}\, Y_i^2$ be the variance of the i-th component, i.e. transform coefficient. Note that we actually mean a random variable when we talk about coefficients/components. If we use N scalar quantizers to quantize $\boldsymbol{Y}$, the optimal high-rate bit allocation is easily found using Lagrangian optimization.[1] We get an average distortion of the form $D = C(\prod_{i=1}^{N} \sigma_i^2)^{1/N} e^{-2R}$, with C a constant. This can be compared with the distortion of a scalar quantizer applied to the X_i's, which is $D = C\sigma^2 e^{-2R} = C[\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, X_i^2]e^{-2R}$. In fact, the *transform coding gain* is defined as the ratio of the distortion of direct scalar quantization of the signal samples over scalar quantization of the transform coefficients (with bit allocation):

$$G_{TC} = \frac{\frac{1}{N}\sum_{i=1}^{N} \sigma_i^2}{\left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N}} = \frac{\mathfrak{A}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)}{\mathfrak{G}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)}. \tag{29}$$

In purely algebraic terms, equation (29) is the ratio of the arithmetic mean $\mathfrak{A}$ of the coefficient variances to their geometric mean $\mathfrak{G}$, which is often used as the "axiomatic" definition of coding gain (the notation $\mathfrak{A}, \mathfrak{G}$ is from [19]). Our short derivation gives some additional insight into the implicit assumptions, namely high rate and (near-)Gaussianity.

Now it is obvious that we can define a measure of coding gain for magnitude classifying quantization by considering the ratio of the Gaussian upper bound to the high-rate upper bound (22).

Definition 2 *The coding gain for optimal*[2] *magnitude classifying quantization is*

$$G_{MCQ} = \frac{c(0)}{c(t_0)} = \frac{\sigma^2}{c(t_0)\sigma^2} = \frac{\mu(t_0)\,\sigma_1^2(t_0) + (1-\mu(t_0))\,\sigma_0^2(t_0)}{e^{2h_b(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}, \tag{30}$$

where t_0 is the threshold yielding the tightest upper bound in Theorem 3.

[1] This uses the assumption that either the signal is a correlated Gaussian process (then any orthonormal transform will yield Gaussian coefficients), or at least that the signal components X_i and the transform coefficients Y_i have the same "marginal" high-rate $D(R)$ behavior of the form $D = Ce^{-2R}$.

[2] Here *optimal* refers to the tightest upper bound of Theorem 3; directly optimizing a MCQ would yield tighter bounds, because significant and insignificant samples differ in $D(R)$ behavior.

Except for the additional side information term $e^{2h_b(\mu(t_0))}$, this definition corresponds to the classical coding gain (29) for two sources with weights $\mu(t_0)$ and $1-\mu(t_0)$. This similarity opens a new perspective on transform coding: instead of considering each transform coefficient as a distinct random variable, we mix all coefficients together and use a quantizer for the marginal density. A transform that has high classical coding gain will have a peaked marginal density, so that the MCQ coding gain will also be large. At the same time, the mixing approach obviously entails a loss in coding gain, which we study by means of an example.

Example 1 (Coding Gain Loss for Gaussian Mixtures) If the transform outputs zero mean Gaussian coefficients, where each has one of just two distinct variances, the resulting marginal density will be a two-component Gaussian mixture like the one studied in Section IV. We get the largest classical coding gain if for every sample we know from which of the two sources it came from. That situation corresponds exactly to the oracle lower bound presented in Section IV-A, and the coding gain is simply the distance in dB to the Gaussian upper bound. The coding gain loss is the ratio of MCQ coding gain (30) to classical coding gain (29), or the distance in dB from the lower bound (27) to the high-rate upper bound (22):

$$\Delta_{CG} = \frac{e^{2h_b(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}{\sigma_{m0}^{2(1-w_1)}\sigma_{m1}^{2w_1}}.$$

Note that here $\sigma_0^2(t_0)$ denotes the variance of the sub-threshold samples, while σ_{m0}^2 is the first mixture variance. Figure 6 contains contour plots of (a) the coding gain and (b) the coding gain loss Δ_{CG} for different ratios $\theta^2 = \sigma_{m1}^2/\sigma_{m0}^2$ of the mixture variances and weights $w_1 = 1 - w_0$ ($\theta = 1$ is the Gaussian pdf). Large θ and small w_1 lead to peaked densities; for example the wavelet coefficient mixture from Section IV has $\theta \approx 30.9$ and $w_1 \approx 0.09$. From the graph, we see that these values correspond to a loss of about 2.5 dB, which can be verified by checking the distance between the high-rate bounds in Figure 5.

The above definition of coding gain loss is based on the assumption that we are actually mixing two Gaussian sources with distinct variances (i.e. $\theta > 1$). What if we only have a single source with the same marginal mixture density? Then the lower bound is not achievable for $\theta > 1$ and thus a better definition of coding gain loss is the ratio of the high-rate upper bound

to the Shannon lower bound:

$$\Delta_{CG(SLB)} = \frac{e^{2h_b(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}{\exp[2h(X) - \log(2\pi e)]}.$$

The differential entropy $h(X)$ has to be computed with numerical integration methods. Figure 7 plots the coding gain (see Definition 4 in Section V-C) and the coding gain loss $\Delta_{CG(SLB)}$ for this case. The loss is remarkably low over a wide range of parameter values, which shows that the magnitude classification quantization approach is very effective for such sources. Let us also remark that in this example the optimal MCQ threshold t_0 was always larger than the threshold for the maximum likelihood classification, $t_{ML} = \sqrt{\log\theta^2/(1-\theta^{-2})}\sigma_{m0}$. This is quite natural, since the goal of the classification is a tight distortion bound, not the optimal distinction of the two component sources.

V. A Measure of Sparseness

In this section we will argue that the geometric mean $\mathfrak{G}(|X|) = \exp(\mathrm{E}\log|X|)$ of a scalar random variable X, respectively its logarithm, is a useful single-letter measure of sparseness. Under the condition that the distribution $F_X(x)$ is continuous at zero, i.e. that $\Pr\{X=0\}=0$, the geometric mean is well defined and clearly measures sparseness. The more probability mass is concentrated around zero, the smaller $\mathfrak{G}(|X|)$ will be and the sparser a vector of samples of X will look.

We will show that in combination with the variance, the geometric mean allows us to bound the source entropy and therefore characterize the high-rate $R(D)$ behavior of sparse sources. The latter fact follows from the tightness of the Shannon lower bound, that is $R(D) - R_{SLB}(D) \to 0$ as $D \to 0$, see e.g. [5, Sec. 4.3.4]. To start, we prove that the geometric mean provides an upper bound on the MCQ compression gain.

Definition 3 *The* normalized squared geometric mean *of a memoryless source with finite variance is defined as*

$$M_G(X) = \frac{\exp(\mathrm{E}\log X^2)}{\mathrm{E}\,X^2} = \frac{\mathfrak{G}(X^2)}{\mathfrak{A}(X^2)} \tag{31}$$

By the arithmetic-geometric mean inequality we have $M_G \leq 1$, with equality iff the source magnitude is constant ($|X| = \sigma$).

Theorem 4 *The factor $c(t)$ in the high-rate bound (Theorem 3) is lower bounded by the normalized squared geometric mean:*

$$c(t) \geq M_G(X) \tag{32}$$

Proof: We bound $\mathrm{E}\log X^2$ by applying Jensen's inequality to the significant and insignificant samples separately:

$$\begin{aligned}
\mathrm{E}\log X^2 &= \Pr\{X^2 < t^2\}\,\mathrm{E}[\log X^2 | X^2 < t^2] + \Pr\{X^2 \geq t^2\}\,\mathrm{E}[\log X^2 | X^2 \geq t^2] \\
&\leq [1-\mu(t)] \log \mathrm{E}[X^2 | X^2 < t^2] + \mu(t) \log \mathrm{E}[X^2 | X^2 \geq t^2] \\
&= [1-\mu(t)] \log \tfrac{1-A(t)/\sigma^2}{1-\mu(t)} + \mu(t) \log \tfrac{A(t)/\sigma^2}{\mu(t)} + \log \sigma^2.
\end{aligned}$$

Now subtract $\log\sigma^2$ from both sides and observe that $h_b(\mu(t)) \geq 0$. Exponentiating both sides proves the theorem. ■

An immediate consequence is that $1/M_G$ is an upper bound to the MCQ coding gain G_{MCQ} (30). If $M_G \ll 1$, there *may* exist a t_0 such that $c(t_0) \ll 1$, i.e. such that there is a large coding gain. On the other hand, if M_G is closer to 1, the coding gain is necessarily small.

We could also call $G_s = M_G^{-1}$ the *sample coding gain* for the following reason. Consider a block of n i.i.d. samples from a memoryless source with a continuous distribution (in particular with $\Pr\{X=0\}=0$), such that Fubini's theorem applies to the product density. The squared geometric mean of these n samples is $\mathfrak{G}_n^2(\boldsymbol{x}) = (\prod_{i=1}^n x_i^2)^{1/n}$, while its expected value is

$$\mathrm{E}\,\mathfrak{G}_n^2(\boldsymbol{X}) = \int \prod_{i=1}^n |x_i|^{2/n} \prod_{i=1}^n f(x_i)\,\mathrm{d}\boldsymbol{x} = \prod_{i=1}^n \int |x_i|^{2/n} f(x_i)\,\mathrm{d}x_i = (\mathrm{E}\,|X|^{2/n})^n.$$

If we let the block size go to infinity, we obtain the geometric mean of the source [19, p. 139]:

$$\mathfrak{G}(X^2) = \lim_{n\to\infty} \left(\mathrm{E}\,|X|^{2/n}\right)^n = \lim_{p\to 0+} \left(\mathrm{E}\,|X|^{2p}\right)^{1/p} = \exp(2\,\mathrm{E}\log|X|). \tag{33}$$

The same follows for the arithmetic mean and hence the sample coding gain G_s deserves its name.

At this point we want to remark that the quasi-norm $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ with $0 < p \leq 1$ is often used as a sparseness measure, see for example [15] and references therein. The obvious question is: how to choose p? If $\boldsymbol{x}$ is a sample of i.i.d. random variables, the choice $p = 1/n$ will yield the geometric mean as $n \to \infty$, by equation (33). This is a strong argument in favor

of the geometric mean as a sparseness measure for random variables. In this respect, it is also interesting to observe that $\lim_{p\to 0+} \|\boldsymbol{x}\|_p^p$ is equal to the Hamming weight $w_H(\boldsymbol{x})$, that is the "strictest" sparseness measure in the sense that only values that are exactly zero contribute to sparseness.

A. Lower Bound on Differential Entropy

The logarithm of the geometric mean, $\mathrm{E}\log|X|$, yields a lower bound on the entropy of continuous random variables with one- or two-sided monotone densities. In turn, this can be used to bound high-rate $R(D)$. We first prove a weaker bound that has the appeal of displaying the relationship with an analogous bound for discrete entropy. Then we will prove a bound which is tight for the class of monotone densities considered.

Proposition 2 *Let X be a finite variance random variable with a monotone one-sided pdf f and range $[x_0, \infty)$ or $(-\infty, x_0]$. Then*

$$h(X) \geq \mathrm{E}\log|X - x_0|. \tag{34}$$

Proof: Without loss of generality, consider a pdf f which is monotone decreasing on $[x_0, \infty)$. The monotonicity implies that f is Riemann-integrable, and the finite variance ensures that the entropy integral is finite (by the Gaussian upper bound). We will approximate the integral $h(X) - \mathrm{E}\log|X - x_0| = -\int_{x_0}^{\infty} f(x)\log(|x - x_0|f(x))\,\mathrm{d}x$ by a Riemann sum with step size Δ. Let $x_i = x_0 + i * \Delta$ and $p_i = f(x_i)\Delta$, for $i = 1, 2, \ldots$. By monotonicity, we have $p_1 \geq p_2 \geq \ldots$ and hence

$$1 \geq \sum_{i=1}^{\infty} p_i \geq \sum_{i=1}^{n} p_i \geq n p_n. \tag{35}$$

Thus we can write

$$\begin{aligned} h(X) - \mathrm{E}\log|X - x_0| &= \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log(|x_n - x_0|f(x_n)) \\ &= \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log\left(n\Delta \cdot \frac{p_n}{\Delta}\right) \\ &\geq \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log(1) = 0, \end{aligned} \tag{36}$$

where the inequality follows from taking the logarithm of (35). ■

Remark: The inequality (35) was used by Wyner to prove an analogous bound for discrete entropy [20].

By using a different proof technique, we obtain a stronger result:

Theorem 5 *Let X be a finite variance random variable with a monotone one-sided pdf f and range $[x_0, \infty)$ or $(-\infty, x_0]$. Then*

$$h(X) \geq \mathrm{E} \log |X - x_0| + 1, \tag{37}$$

with equality iff f is a uniform density.

Proof: For simplicity we assume f to be decreasing on $[0, \infty)$. Let $\mathcal{B}$ be the set of all such monotone decreasing, finite variance pdf's on $[0, \infty)$. It is easy to verify that $\mathcal{B}$ is a convex set. Its boundary is given by the set of all finite variance uniform densities:

$$\partial\mathcal{B} = \{u(a, x) : a \in (0, \infty)\}, \tag{38}$$

where

$$u(a, x) = \begin{cases} 1/a & \text{if } 0 \leq x \leq a, \\ 0 & \text{else.} \end{cases} \tag{39}$$

Too see that (38) is indeed the boundary of $\mathcal{B}$, observe first that no uniform density $u(a, x)$ can be written as a nontrivial convex combination of two distinct monotone decreasing densities. Moreover, any $f \in \mathcal{B}$ can be written as a convex combination of elements of $\partial\mathcal{B}$:

$$f(x) = \int_0^\infty \lambda(a) u(a, x) \, \mathrm{d}a, \tag{40}$$

where $\lambda(a) = -af'(a)$, as can be shown with some simple calculus. $\lambda(a)$ is a proper distribution if $f(x)$ has finite variance (in particular, $\lim_{x\to\infty} xf(x) = 0$) and if $f'(x) \leq 0$, which is indeed the case for monotone decreasing f. Using the standard extensions to distributions, (40) also holds if f contains a countable number of steps, e.g. if it is piecewise constant. In fact, (40) is nothing but a disguised version of the "layer cake representation" of f, namely $f(x) = \int_0^\infty \chi_{\{f>t\}}(x) \, \mathrm{d}t$. This follows from the monotonicity of f.

Looking at (37), we see that

$$h(X) - \mathrm{E} \log X = -\int_0^\infty f(x) \log(xf(x)) \, \mathrm{d}x \tag{41}$$

is a concave-$\cap$ functional of f, since $h(X)$ is concave and $\mathrm{E}\log X$ is linear in f. Therefore a minimum of (41) over the convex set $\mathcal{B}$ must necessarily lie on its boundary $\partial\mathcal{B}$. We insert an arbitrary boundary element $u(a,x)$ $(0 < a < \infty)$ in (41) to obtain

$$\begin{aligned} h(X) - \mathrm{E}\log X &= -\int_0^\infty u(a,x)\log(xu(a,x))\,\mathrm{d}x \\ &= -\int_0^a \tfrac{1}{a}\log\tfrac{x}{a}\,\mathrm{d}x \\ &= \log a - \tfrac{x}{a}(\log x - 1)\big|_0^a \\ &= 1. \end{aligned} \tag{42}$$

Since (42) holds for any a, we conclude that it is the global minimum, thus proving (37) and one part of the "iff". To prove the other part, it suffices to observe that $h(X) - \mathrm{E}\log X$ is a strictly concave functional and thus will be larger than (42) in the interior $\mathcal{B} \setminus \partial\mathcal{B}$. ■

We define a *weakly unimodal* density with mode x_0 to be a pdf which is monotone increasing (non-decreasing) on $(-\infty, x_0]$ and monotone decreasing (non-increasing) on $[x_0, \infty)$.

Corollary 6 *Let X be a finite variance random variable with weakly unimodal pdf f such that $\Pr\{X \le x_0\} = \alpha$, where x_0 is the mode. Then*

$$h(X) \ge \mathrm{E}\log|X - x_0| + 1 + h_b(\alpha). \tag{43}$$

For a density that is symmetric about x_0, $f(-x - x_0) = f(x - x_0)$, this reduces to

$$h(X) \ge \mathrm{E}\log|X - x_0| + 1 + \log 2. \tag{44}$$

Proof: We view the weakly unimodal pdf f as a mixture of two non-overlapping monotone one-sided densities, $f_l(x)$ and $f_r(x)$, with weights α and $1-\alpha$, respectively. Without loss of generality we can assume $x_0 = 0$. Then,

$$\begin{aligned} h(X) - \mathrm{E}\log|X| &= -\mathrm{E}_f \log[|X|f(X)] \\ &= -\int_{-\infty}^0 \alpha f_l(x)\log(-x\alpha f_l(x)) - \int_0^\infty (1-\alpha) f_r(x)\log(x(1-\alpha)f_r(x)) \\ &= h_b(\alpha) - \alpha\,\mathrm{E}_{f_l}\log[|X|f(X)] - (1-\alpha)\,\mathrm{E}_{f_r}\log[|X|f(X)] \\ &\ge h_b(\alpha) + 1, \end{aligned} \tag{45}$$

where the last inequality follows from Theorem 5. ■

B. Upper Bound on Differential Entropy

If both the variance and the geometric mean are known, an upper bound on the entropy can be easily obtained via the maximum entropy approach. Owing to the assumptions made in this variational approach, the results in this section hold for random variables which have an absolutely continuous distribution function $F(x)$ with probability density $f(x) = F'(x)$.

Theorem 7 *The two-sided maximum entropy pdf given the constraints* $\mathrm{E}\, X^2 = \sigma^2$ *and* $\mathrm{E} \log |X| = \theta$ *is*

$$f(x) = \Gamma^{-1}(\tfrac{u}{2}) \left(\tfrac{u}{2\sigma^2}\right)^{u/2} |x|^{u-1} \exp\left(-\tfrac{ux^2}{2\sigma^2}\right). \tag{46}$$

The shape parameter $u > 0$ *is obtained by solving*

$$\mathrm{E} \log |X| = \tfrac{1}{2}\Psi(\tfrac{u}{2}) - \tfrac{1}{2} \log \tfrac{u}{2\sigma^2} \stackrel{!}{=} \theta. \tag{47}$$

For any $\theta \leq \log \sigma$ *there is a unique solution, since (47) is strictly monotone increasing in* u. *The resulting entropy is*

$$h(\sigma, \theta) = \tfrac{u}{2} - \tfrac{u-1}{2}\Psi(\tfrac{u}{2}) + \log \Gamma(\tfrac{u}{2}) - \tfrac{1}{2} \log \tfrac{u}{2\sigma^2} \tag{48}$$

Setting $u = 1$ *yields the Gaussian density and thus the global entropy maximum given the variance constraint alone.*

Corollary 8 *The entropy of any random variable with probability density* f *satisfying* $\mathrm{E}\, X^2 = \sigma^2$ *and* $\mathrm{E} \log |X| = \theta$ *is upper bounded by (48).*

Proof: Before proving Theorem 7 we state and prove an auxiliary lemma on one-sided densities.

Lemma 1 *The maximum entropy pdf on* $[0, \infty)$ *given the constraints* $\mathrm{E}\, X^2 = \sigma^2$ *and* $\mathrm{E} \log X = \theta$ *is*

$$g(x) = 2\Gamma^{-1}(\tfrac{u}{2}) \left(\tfrac{u}{2\sigma^2}\right)^{u/2} |x|^{u-1} \exp\left(-\tfrac{ux^2}{2\sigma^2}\right), \tag{49}$$

with shape parameter $u > 0$ *obtained by solving (47).*

The lemma can be derived using the calculus of variations [3, Chap. 11]. The constraints yield the following functional:

$$J(f) := \int_0^\infty f(x)[-\log f(x) + \lambda_1 + \lambda_2 x^2 + \lambda_3 \log x]\,\mathrm{d}x.$$

"Differentiating" with respect to f and setting the resulting expression to zero shows that the maximizing density has the form

$$f(x) = e^{\lambda_1 - 1} x^{\lambda_3} e^{\lambda_2 x^2}.$$

The three constraints are satisfied by $\lambda_3 = u - 1$, $\lambda_2 = -u/2\sigma^2$ and $e^{\lambda_1 - 1} = 2(-\lambda_2)^{u/2}/\Gamma(u/2)$.

We need to show that $\mathrm{E}\log|X|$ is monotone increasing, so that the mapping between θ and u is one-to-one. By Jensen's inequality we have $\mathrm{E}\log|X| \le \frac{1}{2}\log \mathrm{E}\,X^2 = \log\sigma$. Let $v = u/2$ and $\phi(v) = 2(\mathrm{E}\log|X| - \frac{1}{2}\log \mathrm{E}\,X^2) = \psi(v) - \log v$. Using a standard integral representation for $\psi(v)$ [21] we obtain

$$\phi(v) = -\frac{1}{2v} - 2\int_0^\infty \frac{t\,\mathrm{d}t}{(t^2+v^2)(e^{2\pi t}-1)} \qquad v > 0. \tag{50}$$

The first derivative,

$$\phi'(v) = -\frac{1}{2v^2} + 4\int_0^\infty \frac{vt\,\mathrm{d}t}{(t^2+v^2)^2(e^{2\pi t}-1)}, \tag{51}$$

is strictly positive for $v > 0$, so $\mathrm{E}\log|X|$ is indeed monotone increasing. By bounding the integral in (50) one can further show that $\lim_{u\to\infty} \mathrm{E}\log|X| = \log\sigma$. This proves the lemma.

For a two-sided random variable X with pdf f, the lemma implies that the pdf of the magnitude $|X|$ must be of the form (49), or $f(-x) + f(x) = g(|x|)$. Furthermore, the entropy cannot be maximal unless $f(-x) = \mu g(x)$ and $f(x) = (1-\mu)g(x)$ for $x \ge 0$ and some $0 \le \mu \le 1$. Now, if we write the entropy integral as $-\int_0^\infty f(-x)\log f(-x)\,\mathrm{d}x - \int_0^\infty f(x)\log f(x)\,\mathrm{d}x$, we see immediately that this is maximal iff $\mu = 1/2$, that is $f(-x) = f(x) = g(x)/2$. This proves the theorem; the corollary is implicit in the maximum entropy approach. ■

Theorem 9 *The maximum entropy (48) for a finite variance σ^2 has the following asymptotic behavior as the geometric mean* $\exp(\theta)$ *goes to zero, resp. $\theta \to -\infty$:*

$$h(\sigma,\theta) \simeq \theta + \log(-2e\theta), \quad \theta \to -\infty \tag{52}$$

Proof: Note that $\theta \to -\infty$ corresponds to $u \to 0+$. Let

$$\begin{aligned}\Delta &= h(\sigma,\theta) - \theta - \log(-2e\theta) \\ &= \frac{u}{2} - \frac{u}{2}\Psi\left(\frac{u}{2}\right) - 1 + \log\left(\frac{\Gamma(\frac{u}{2})}{-\Psi(\frac{u}{2}) + \log\frac{u}{2\sigma^2}}\right). \end{aligned} \tag{53}$$

To prove $\lim_{u\to 0+}\Delta = 0$, which is slightly stronger than required, we use the functional relationships $\Gamma(x+1) = x\Gamma(x)$, $\Psi(x+1) = \Psi(x) + \frac{1}{x}$ and the truncated series expansions $\Gamma(x+1) = 1 - \gamma x + o(x^2)$, $\Psi(x+1) = -\gamma + \frac{\pi^2}{6}x + o(x^2)$, both for $|x| < 1$ (see e.g. [21]; γ is Euler's constant). We have

$$\lim_{u\to 0+} \tfrac{u}{2}\Psi(\tfrac{u}{2}) = \lim_{u\to 0+}[-1 - \gamma\tfrac{u}{2} + \tfrac{\pi^2}{24}u^2 + o(u^3)] = -1,$$

hence $\lim_{u\to 0+}\Delta$ is equal to the limit of the logarithm in (53). But

$$\lim_{u\to 0+} \frac{\Gamma(\frac{u}{2})}{-\Psi(\frac{u}{2}) + \log\frac{u}{2\sigma^2}} = \lim_{u\to 0+} \frac{\frac{2}{u}(1 - \frac{\gamma}{2}u + o(u^2))}{\frac{2}{u} + \log\frac{u}{2\sigma^2} + \gamma - \frac{\pi^2}{12}u + o(u^2)} = 1.$$

This can be easily seen by extending the fraction by $\frac{u}{2}$ and observing that $\lim_{u\to 0+} u\log u = 0$. By putting these steps together we obtain $\lim_{u\to 0+}\Delta = 0$. ■

Figure 8 shows the lower bound (44) and the upper bound (48) as a function of $\theta = \mathrm{E}\log|X|$ for unit-variance random variables with symmetric unimodal densities. The global maximum of the upper bound corresponds to the unit-variance Gaussian density, which has $\mathrm{E}\ln|X| \approx -0.635$. As a consequence of Theorem 9, the gap between the lower and upper bounds is asymptotically equal to $\log\theta$. Also shown is a tightened lower bound for Gaussian mixtures, namely equation (57) below. The crossing between upper and lower bounds is only a seeming contradiction, because in fact it simply means that to the right of the crossing there exist no unimodal densities satisfying both the geometric mean and variance constraints.

C. Mixture versus Vector Coding Gain

For an application example of the geometric mean, we take another look at the Gaussian mixtures in Example 1. There we compared the magnitude classification upper bound for the mixture of two zero mean Gaussians with the oracle lower bound and the Shannon lower bound. Here we will show that there is a simple relationship between the geometric mean of the variances of N Gaussians and the geometric mean of their uniform mixture. This can be used to bound the coding gain of a Gaussian mixture vs. the coding gain for the unmixed sources.

The logarithmic geometric mean of a mixture of N zero mean Gaussians with variances σ_i^2 and mixing weights w_i is

$$\theta = \mathrm{E}\, log|X| = \tfrac{1}{2}\sum_{i=1}^{N} w_i \log \sigma_i^2 - \tfrac{1}{2}\log 2 - \tfrac{1}{2}\gamma, \tag{54}$$

where $\gamma = 0.5772156649\ldots$ is Euler's constant. Comparing this with the *vector* coding gain of N-dimensional Gaussian transform coding (29), setting $w_i = 1/N$ (uniform mixture) and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$ we see that

$$\sum_{i=1}^{N} w_i \log \sigma_i^2 = \log(\sigma^2/G_{TC}). \tag{55}$$

This is the desired relationship between the coding gain for bit allocation over N independent Gaussian sources and the geometric mean of the mixture of these sources.

At this point we explicit a definition that has already been used in Example 1.

Definition 4 *The coding gain for an i.i.d. (scalar) source is defined as the ratio of the Gaussian $D(R)$ upper bound to the Shannon lower bound:*

$$G_{SLB} = \frac{2\pi e\sigma^2}{\exp(2h(X))}. \tag{56}$$

It measures the coding gain achieved by using a codebook matched to the source instead of a Gaussian codebook.

Via the geometric mean we can bound the mixture entropy $h(X)$ and from that the mixture coding gain G_{SLB} (56). Therefore the upper bound of Corollary 8 leads to a lower bound on G_{SLB}. In the same manner we could use Corollary 6 to obtain an upper bound. However, this can be tightened by the same approach as in Section IV-A, namely by lower bounding the Gaussian mixture entropy by conditioning on the hidden state selecting the mixture components:

$$h(X) \geq h(X|S) = \tfrac{1}{2}\sum w_i \log(2\pi e\sigma_i^2). \tag{57}$$

In combination with (55) and (56) this yields $G_{SLB} \leq G_{TC}$, which simply means that mixing does not necessarily inflict a performance penalty. Figure 9 is a plot of the upper and lower bounds for mixture vs. vector coding gain.

What exactly are we comparing? On the one hand, we have the classical vector coding gain for N independent Gaussian sources. On the other hand, the coding gain for a mixture source

that outputs one of these N sources uniformly at random. As an example, consider a transform that outputs N independent zero mean Gaussian components. If we know the variance of each component, like e.g. in the KLT case, we can achieve the vector (transform) coding gain. If however only the distribution of the variances is known, then we can design a codebook for the corresponding scalar mixture source and still achieve the mixture coding gain. This is the case of transforms with "known eigenvalue distribution", but "unknown positions". Intuitively, wavelet transforms lie between these two extremes, since e.g. coefficient variances are correlated across scales (however this also violates the underlying independence assumption). In summary, the lower curve in Figure 9 bounds the maximum performance loss of a "naive" one-dimensional system compared to one with perfect side information.

Appendix

Definition 5 (Rate distortion function of a DMS) *Let $X \sim P$ be a discrete memoryless random variable, $\rho(x, \hat{x})$ a single-letter distortion measure, $Q_{\widehat{X}|X}(k|j)$ a conditional distribution (defining a random codebook), and $P_{X,\widehat{X}}(j,k) = P(j)Q(k|j)$ the corresponding joint distribution. The* average distortion *associated with $Q(k|j)$ is*

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\rho(j,k). \tag{58}$$

If a conditional probability assignment satisfies $d(Q) \leq D$ it is called D-admissible. *The set of all D-admissible Q is $Q_D = \{Q(k|j) : d(Q) \leq D\}$. The average mutual information ("description rate") induced by Q is*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \frac{Q(k|j)}{Q(k)}, \tag{59}$$

where $Q(k) = \sum_j P(j)Q(k|j)$. The rate distortion function $R(D)$ is defined as

$$R(D) = \min_{Q \in Q_D} I(Q)$$

This convex optimization problem can be solved with the method of Lagrange multipliers [5], [3, Section 13.7]. We start with the functional

$$J(Q) = I(Q) + \lambda d(Q) + \sum_j \nu_j \sum_k Q(k|j),$$

where the last term comes from the constraint that $Q(k|j)$ is a proper conditional distribution, i.e. satisfies $\sum_k Q(k|j) = 1$. The minimizing conditional distribution can be computed as

$$Q(k|j) = \frac{Q(k)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}}. \tag{60}$$

The marginal $Q(k)$ has to satisfy the following $\widehat{N} = |\widehat{\mathcal{X}}|$ conditions:

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} = 1 \qquad \text{if } Q(k) > 0, \tag{61}$$

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} \leq 1 \qquad \text{if } Q(k) = 0. \tag{62}$$

Inequality (62) stems from the Kuhn-Tucker conditions (for a detailed derivation of the above see Section 13.7 in [3]). The solution of the problem is further simplified by the following theorem by Berger:

Theorem 10 *[5, Theorem 2.6.1] No more than N reproducing letters need be used to obtain any point on the $R(D)$ curve that does not lie on a straight-line segment. At most, $\widehat{N} = N + 1$ reproducing letters are needed for a point that lies on a straight-line segment.*

Acknowledgments

The authors wish to thank Emre Telatar for helpful discussions.

References

[1] Stéphane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, 1999.

[2] Stéphane Mallat and Frédéric Falzon, "Analysis of low bit rate image transform coding," *IEEE Trans. Signal Proc.*, vol. 46, pp. 1027–1042, April 1998.

[3] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[4] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc., Special Issue on Wavelets and Signal Processing*, vol. 41, no. 12, pp. 3445–3462, December 1993.

[5] Toby Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, 1971.

[6] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[7] Claudio Weidmann and Martin Vetterli, "Rate distortion behavior of threshold-based nonlinear approximations," in *Proc. Data Compression Conference*, Snowbird, Utah, March 2000, pp. 333–342.

[8] Claudio Weidmann, *Oligoquantization in Low-Rate Lossy Source Coding*, Ph.D. thesis, Swiss Federal Institute of Technology, Lausanne, July 2000.

[9] David J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 301–309, May 1975.

[10] Herbert Gish and John N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, September 1968.

[11] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day, New York, 1964.

[12] Hanan Rosenthal and Jacob Binia, "On the epsilon entropy of mixed random variables," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1110–1114, September 1988.

[13] András György, Tamás Linder, and Kenneth Zeger, "On the rate-distortion function of random vectors and stationary sources with mixed distributions," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 2110–2115, September 1999.

[14] David L. Donoho, Martin Vetterli, R.A. DeVore, and Ingrid Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 2435–2476, October 1998.

[15] N. Saito, B. M. Larson, and B. Bénichou, "Sparsity vs. statistical independence from a best-basis viewpoint," in *Wavelet Applications in Signal and Image Processing VIII*, 2000, Proc.SPIE 4119, pp. 474–486, `http://www.math.ucdavis.edu/~saito/publications/`.

[16] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Trans. Signal Proc.*, vol. 46, pp. 886–902, April 1998.

[17] Are Hjørungnes and John M. Lervik, "Jointly optimal classification and uniform threshold quantization in entropy constrained subband image coding," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, 1997, vol. 4, pp. 3109–3112.

[18] Robert M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, July 1973.

[19] G. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, 2nd edition, 1952.

[20] A.D. Wyner, "An upper bound on the entropy series," *Inform. Contr.*, vol. 20, pp. 176–181, 1972.

[21] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, fifth edition, 1994.

[22] Claudio Weidmann and Martin Vetterli, "Rate-distortion analysis of spike processes," in *Proc. Data Compression Conference*, Snowbird, Utah, March 1999, pp. 82–91.

[23] Claudio Weidmann, "Rate distortion bounds via threshold-based classification," in *Proc. IEEE Int. Symp. Information Theory*, Washington DC, June 2001, p. 169.

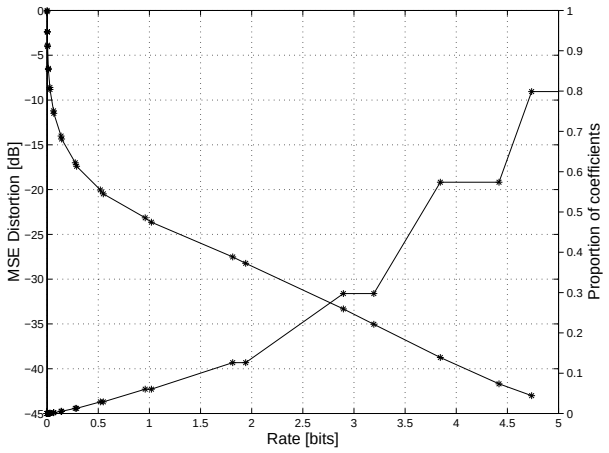


Fig. 1. Typical operational distortion rate curve of a wavelet image coder (decreasing curve, left scale). At low rates, only a small fraction of coefficients is quantized to nonzero values, all the others are not used in the reconstruction of the image (increasing curve, right scale).

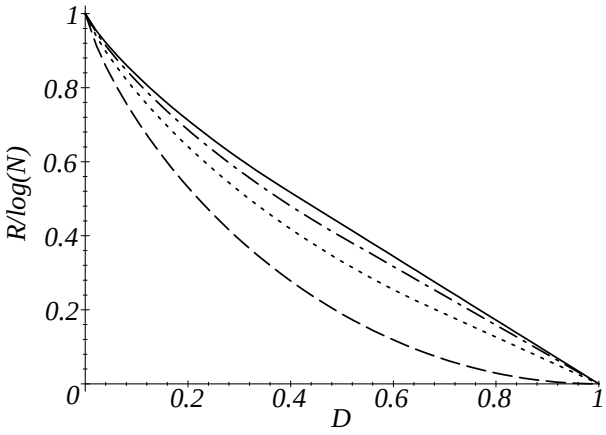


Fig. 2. $R(D)$ for single spike with Hamming distortion, $N = 2$ (bottom) up to $N = 5$ (top curve). The rate has been normalized to $1/\log N$. For $N \to \infty$, $R(D)$ becomes a straight line, see (3).

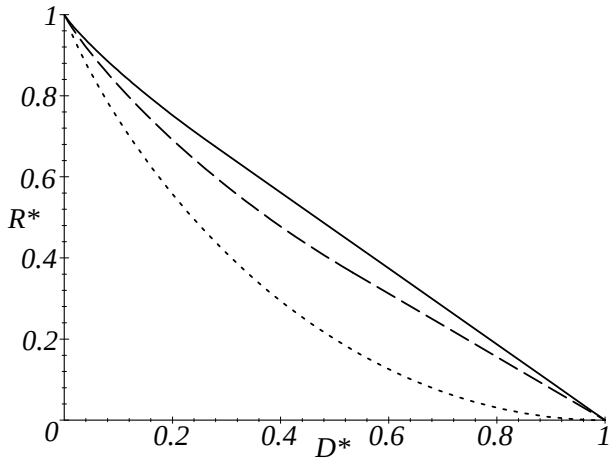


Fig. 3. $R(D)$ for multiple spikes with Hamming distortion, $K = 4, 8, 16$ spikes (top to bottom curve) in $N = 32$ positions. Rate, distortion normalized to $R^* = R/\log\binom{N}{K}$ and $D^* = D/K$.

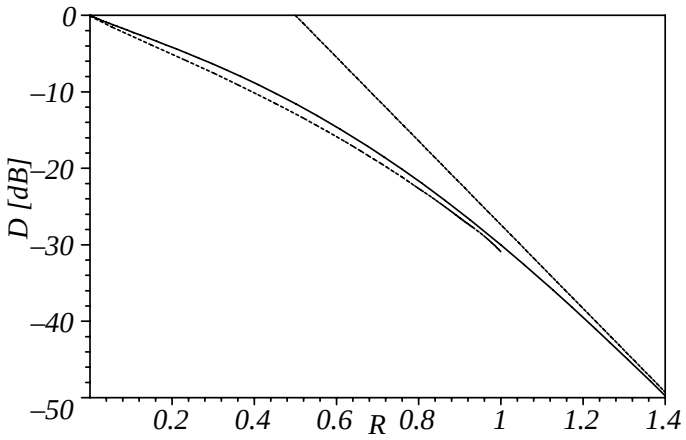


Fig. 4. Bernoulli-Gaussian spike with $p = 0.11$: empirical $D(R)$, upper bound (20) and trivial upper bound $B(0, R)$ (bottom to top curve). Normalized to unit variance.

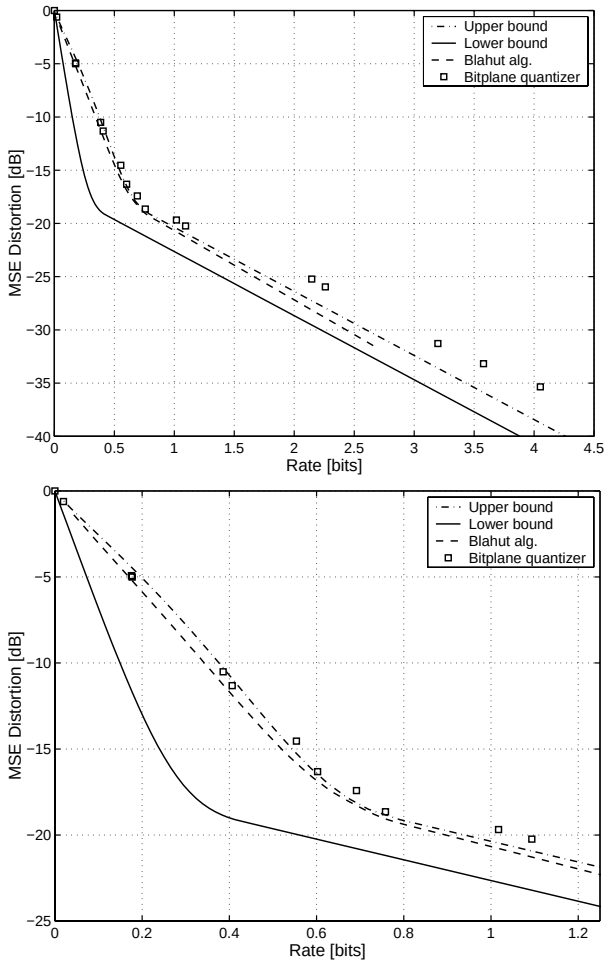
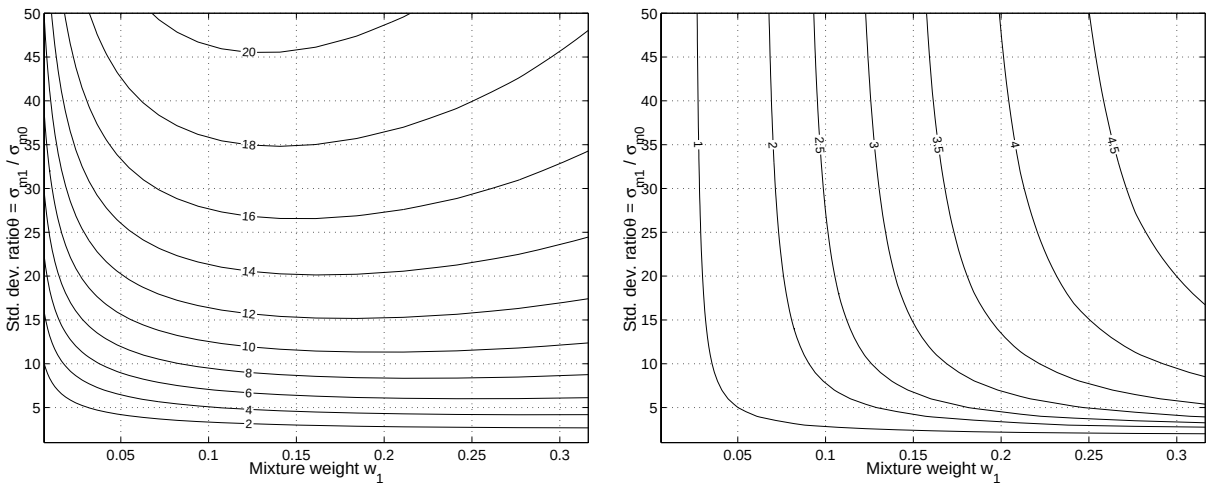
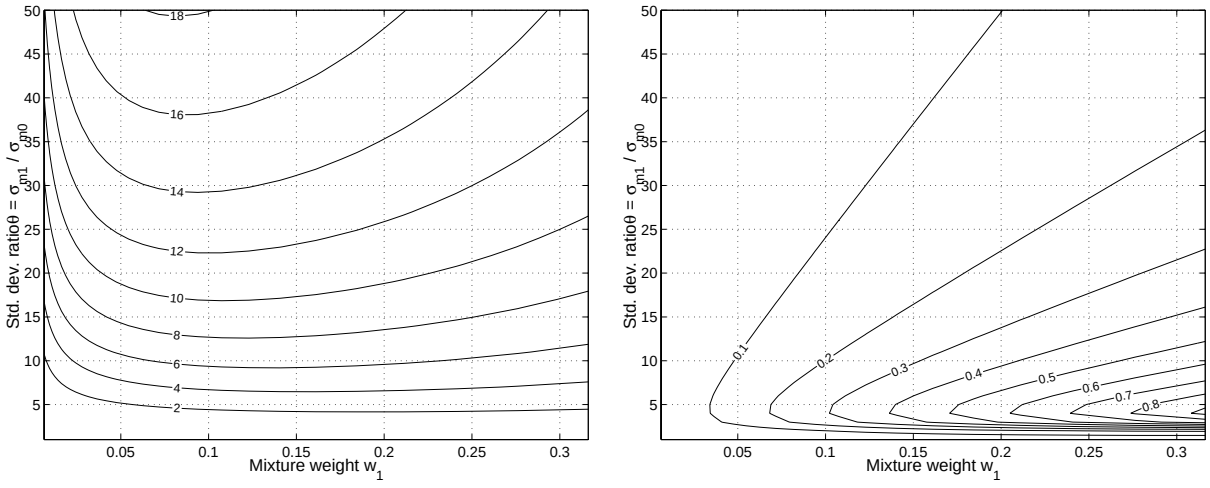


Fig. 5. Gaussian mixture model for wavelet (detail) coefficients: Upper and lower distortion rate bounds for Gaussian mixture model. Model parameters, normalized to unit variance: $p = 0.9141$, $\sigma_1^2 = 0.01207$ and $\sigma_2^2 = 11.51$. The middle curve is the empirical $D(R)$, the boxes denote (R, D) points achieved with a bitplane quantizer. At bottom, detail of low-rate region.



(a) (b)

Fig. 6. Magnitude classifying quantization (MCQ) of two-component Gaussian mixtures (GM). (a) Coding gain G_{TC} for *unmixed*, separate sources (equivalent to GM lower bound). (b) Coding gain loss relative to G_{TC} for MCQ of the mixture.



(a) (b)

Fig. 7. Magnitude classifying quantization (MCQ) of two-component Gaussian mixtures (GM). (a) Coding gain G_{SLB} for mixture source (equivalent to Shannon lower bound). (b) Coding gain loss relative to G_{SLB} for MCQ of the mixture.

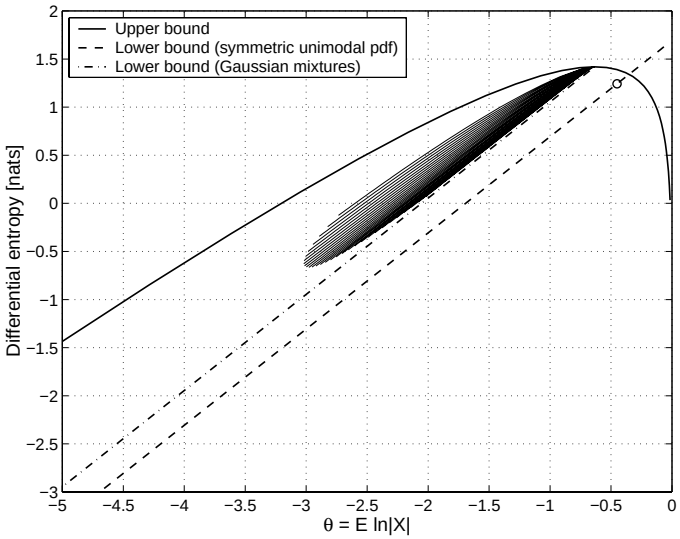


Fig. 8. Differential entropy bounds for symmetric weakly unimodal densities (normalized to unit variance). The "cloud" sweeps the $(\mathrm{E}\ln|X|, h(X))$-pairs of two-component Gaussian mixtures with the same parameter values as in Figures 6 and 7. The circle denotes the unit-variance uniform density for which the lower bound is tight.

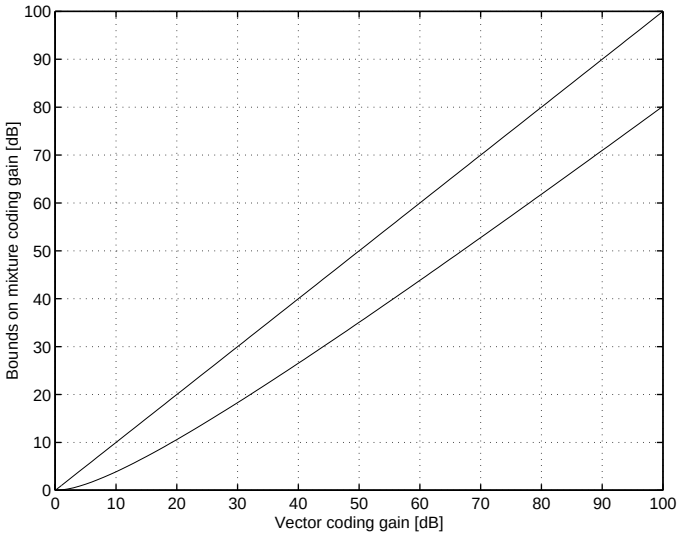


Fig. 9. Bounds for Gaussian mixture vs. vector coding gain.