Rate-Distortion Analysis of Spike Processes

Claudio Weidmann and Martin Vetterli*
Dept. of Electrical Engineering
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
{weidmann,vetterli}@lcav.de.epfl.ch

Abstract

Recent rate distortion analyses of image transform coders are based on a trade-off between the *lossless* coding of coefficient positions vs. the lossy coding of the coefficient values. We propose spike processes as a tool that allows a more fundamental trade-off, namely between *lossy* position coding and lossy value coding. We investigate the Hamming distortion case and give analytic results for single and multiple spikes. We then consider upper bounds for a single Gaussian spike with squared error distortion. The obtained results show a rate distortion behavior which switches from linear at low rates to exponential at high rates.

1 Introduction

The current paradigm for the analysis of low bit rate image transform coding relies essentially on non-linear signal approximation theory [2, 6]. Images are considered to be highly non-Gaussian processes that are not suited for the familiar linear approach, i.e. to (pre)compute coefficient bit allocations based on statistical parameters. Instead, the non-linear approximation tries to find the "best" basis for a given *individual* signal and a given rate. In all practical schemes (wavelet, JPEG) which have been analyzed in this non-linear framework, the "best" basis consists in specifying the indices of a small number of quantized *significant* coefficients which describe the signal with the desired accuracy. This works very well with wavelet transforms of piecewise regular functions, since there will be only a small number of non-zero coefficients, mostly around the signal singularities [4, 7]. The key aspect is the rate trade-off between the *lossless* code for the coefficient positions and the *lossy* code for the values of those coefficients.

By assuming that the signal belongs to certain functional spaces, for which the coefficient decay behavior is known, the rate trade-off for a given proportion of significant coefficients can be computed. This result can be used in turn to get an

*M. Vetterli is also with the Department of EECS, UC Berkeley, Berkeley CA 947020 approximate operational rate distortion curve. We stress the fact that in general this is *not* the information-theoretic rate distortion ($R(D)$) function, which provides a lower bound for the rate necessary to describe a random signal within a given distortion. While operational $R(D)$ is useful to characterize e.g. a family of coding algorithms, it does not tell us how far from the theoretical optimum they operate.

Our goal is to find an alternative, i.e. information-theoretic, framework for the $R(D)$ analysis of such non-linear approximation schemes. As a tool we introduce the *spike process*, which captures the idea of a single isolated non-zero coefficient. This can be extended to multiple spikes by independent superposition or by (more efficient) joint description. We take the simplest possible approach to lossy position coding by using a Hamming metric, i.e. we do not consider the precision with which a position is specified. The next section starts with a rather general definition of spike processes and then investigates the $R(D)$ behavior for Hamming distortion. Along the way we will recall the relevant results from rate distortion theory. Section 3 deals with spikes having Gaussian-distributed values and mean squared error distortion. We finish with first conclusions and some outlooks.

2 Spikes, Hamming Distortion and $R(D)$

2.1 Definitions and single spike case

In the following definition the spike location will be restricted to a finite set. This makes sense, not only because in practice one always deals with a finite number of transform coefficients, but also because a continuous distribution of spike locations would require infinite rate for zero distortion.

Definition 1 *A **spike process** is a memoryless random process which outputs ordered pairs (U, V) of random variables, where the* position *U has a finite alphabet $\mathcal{U}$ with $|\mathcal{U}| = N$, while the* value *V is drawn from an alphabet $\mathcal{V}$ with no particular restrictions.*

We will assume $\mathcal{U} = \{1, 2, \ldots, N\}$, since any other alphabet of size N can always be represented by that set. Further we suppose that $\mathcal{V}$ is a ring. Then a spike process sample (u, v) can be mapped to a $\mathcal{V}^N$-vector simply by $\phi(u, v) = v\mathbf{e}_u$, where $\mathbf{e}_i$ is the i-th basis vector.

Definition 2 *Let $\mathcal{X}$ be the source alphabet, and $\widehat{\mathcal{X}}$ the reconstruction alphabet. A **single-letter distortion measure** (or fidelity criterion) is a non-negative real-valued function $\rho(x, \hat{x}) : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, \infty[$. In the spike case we have $x = (u, v)$, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, and analogously $\hat{x}$, $\widehat{\mathcal{X}}$.*

Probably the simplest distortion measure that can be applied to spike processes is the Hamming distance between the position vectors, i.e. the basis vectors $\mathbf{e}_i$. It can be shown that just one additional reconstruction letter is needed to achieve the rate distortion bound, and it will map to the all-zero vector $\mathbf{0}$. If we define $\widehat{\mathcal{U}} = \{0\} \cup \mathcal{U}$

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:

$$\begin{aligned} \rho(u, v; \hat{u}, \hat{v}) &= w_H(\mathbf{e}_u - \mathbf{e}_{\hat{u}}) \\ &= 2\delta(u - \hat{u})\delta(\hat{u}) + 1 - \delta(\hat{u}) \end{aligned} \tag{1}$$

Thus "giving the right answer" has zero distortion, a wrong answer two, and not answering costs one distortion unit. We proceed with the definition of the (information) rate distortion function for discrete memoryless sources (DMS), closely following the notation in Berger's text [1]. Note that this coincides with the information-theoretic definition ("infimum of achievable rates").

Definition 3 (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*

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\rho(j,k). \tag{2}$$

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*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \frac{Q(k|j)}{Q(k)}, \tag{3}$$

where $Q(k) = \sum_j P(j)Q(k|j)$ and log *is to the base two*. 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 [1],[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

$$Q(k|j) = \frac{Q(k)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}}. \tag{4}$$

*All logs in this paper are to the base 2, unless otherwise specified.

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

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} = 1 \qquad \text{if } Q(k) > 0, \tag{5}$$

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} \leq 1 \qquad \text{if } Q(k) = 0. \tag{6}$$

Inequality (6) 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 (2.6.1 in [1]):

Theorem 1 (Berger) *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.*

As anticipated above, we need just one additional reconstruction letter to achieve $R(D)$. To see that it can only be the all-zero vector, consider the source alphabet $\widehat{\mathcal{U}} \cong \{\mathbf{e}_{\widehat{\mathcal{U}}}\}$, which consists of all vectors of Hamming weight 1. Any other non-zero vector will be at Hamming distance 1 or more from these vectors and thus can only worsen the distortion achieved by the all-zero vector, i.e. exactly 1.

We are now going to derive the rate distortion function for the Hamming criterion (1) in the case of equiprobable positions $P(j) = 1/N$ $(N \geq 2)$, i.e. uniform memoryless distribution of U: $U \sim \mathcal{U}\{1, 2, \ldots, N\}$. The symmetry of the problem suggests the marginal distribution

$$Q = (q_0, q_1 = q_2 = \ldots = q_N = \frac{1 - q_0}{N}). \tag{7}$$

Let us first assume that $q_k > 0$ holds for all k. Then the $N + 1$ conditions (5) have to be met. We make the substitution $\beta = e^{-\lambda}$ and plug our $Q(k)$ into the equation, first for $k \neq 0$:

$$\begin{aligned} \frac{\beta^0}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} + \frac{(N-1)\beta^2}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} &= \frac{1}{P(j)} = N \\ &\vdots \\ q_0((N-1)\beta^2 - N\beta + 1) &= 0. \end{aligned} \tag{8}$$

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

$$\begin{aligned} \frac{N\beta^1}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} &= \frac{1}{P(j)} = N \\ &\vdots \\ (1 - q_0)((N-1)\beta^2 - N\beta + 1) &= 0. \end{aligned} \tag{9}$$

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 plugged into (4) yields

$$Q(k|j) = q_k (N-1)^{1-\rho(j,k)}. \tag{10}$$

Putting (10) into (2) we get the average distortion $d(Q) = 1 - \frac{N-2}{N}(1-q_0)$ and from (3) 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

$$D(R) = 1 - \frac{R}{\log(N-1)} \qquad \text{for } R < \frac{N-2}{N}\log(N-1). \tag{11}$$

When R reaches its upper bound in (11), D reaches $2/N$ and we have $q_0 = 0$. At that point, equation (8) will be satisfied for all β. According to condition (6), equation (9) now becomes an inequality:

$$(N-1)\beta^2 - N\beta + 1 \geq 0. \tag{12}$$

This is satisfied by $\beta \geq 1$ or $\beta \leq \frac{1}{N-1}$, which is equivalent to $\lambda \geq \ln(N-1)$. 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

$$Q(k|j) = \begin{cases} 0, & k = 0 \\ \frac{\beta^{\rho(j,k)}}{1+(N-1)\beta^2}, & k \neq 0 \end{cases} \tag{13}$$

As before we put this into (2) to get $d(Q) = \frac{2(N-1)\beta^2}{1+(N-1)\beta^2}$ and into (3) yielding

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

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. Eliminating β from the last two equations and combining the result with (11) gives the complete rate distortion function of single spike processes with Hamming distortion measure:

$$R(D) = \begin{cases} \log N - \frac{D}{2}\log(N-1) - h\left(\frac{D}{2}\right), & 0 \leq D \leq \frac{2}{N} \\ (1-D)\log(N-1), & \frac{2}{N} < D \leq 1 \end{cases} \tag{14}$$

Figure 1 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 $D(R)$ function of a binary symmetric source (with doubled distortion).

2.2 Generalizing to multiple spikes

The typical output of an image transform is of course not a single spike in one out of N positions, but rather a group of them. Therefore a way to model a transformed picture is to specify K out of N coefficient positions, e.g. those above a given threshold. In the Hamming case this is equivalent to 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 ones 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. Thus the code is geometrically uniform, i.e. the distance (distortion) profile looks the same from any vector of 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

$$w_d = \binom{K}{d}\binom{N-K}{d}, \quad d = 0 \ldots K \tag{15}$$

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

$$\sum_{d=0}^{K} w_d = \sum_{d=0}^{K} \binom{K}{d}\binom{N-K}{d} = \binom{N}{K}. \tag{16}$$

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 the equation (compare with (8, 9))

$$\sum_{d=0}^{K} w_d \beta^{2d} - \binom{N}{K}\beta^K = 0. \tag{17}$$

The solution $\beta = 1$ corresponds again to 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

$$R(D) = (D - K)\log\beta_0, \quad D(\beta_0) < D < K$$

where the bounds on D guarantee $0 < q_0 < 1$ ($D(\beta_0)$ is defined below in (18)). For $q_0 = 0$, any $\beta \leq \beta_0$ will satisfy the Kuhn-Tucker conditions. We define a pseudo-distribution

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

After some calculations, we get a parametric expression for the rate-distortion curve:

$$D(\beta) = \sum_{d=1}^{K} b_d 2d \tag{18}$$

$$R(\beta) = \log\binom{N}{K} + \sum_{d=0}^{K} b_d \log b_d - \sum_{d=0}^{K} b_d \log w_d \tag{19}$$

for $0 < \beta < \beta_0$. The middle term in the expression for R is the negative entropy of our pseudo-distribution b_d (compare with (14)). Figure 1 shows again that for sparse spikes (low K/N) the linear segment dominates the rate distortion behavior.

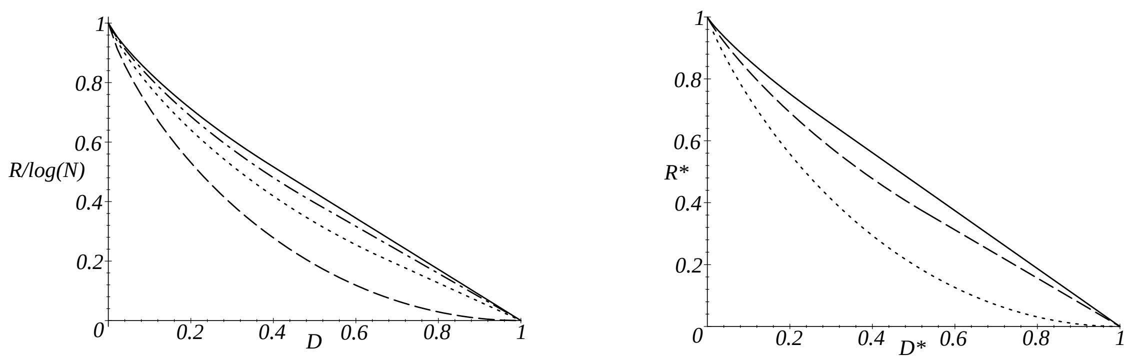

Figure 1: Rate distortion functions for Hamming distortion. *Left:* Single spike with $N = 2$ (bottom) up to $N = 5$ (top curve). The rate has been normalized to $1/log_2 N$. For $N \to \infty$ $R(D)$ becomes a straight line, see (14). *Right:* Multiple spikes ($N = 32$ and $K = 4, 8, 16$, top to bottom). Rate and distortion have been normalized to $R* = R/log_2\binom{N}{K}$ and $D* = D/K$.

3 Single Gaussian Spikes and Squared Error

The coefficients of a wavelet transform are often modeled as having a generalized Gaussian distribution (or approximately Laplacian). Assuming that a R.V. is Gaussian provides an upper bound to its rate distortion function. Hence a *Gaussian spike* with $U \sim \mathcal{U}\{1, 2, \ldots, N\}$ and $V \sim \mathcal{N}(0, \sigma^2)$ is a conservative model for the behavior of wavelet coefficients at signal discontinuities. The appropriate distortion measure is the squared error

$$\begin{aligned} \rho(u, v; \hat{u}, \hat{v}) &= \|v\mathbf{e}_u - \hat{v}\mathbf{e}_{\hat{u}}\|^2 \\ &= \delta(u - \hat{u})(v - \hat{v})^2 + (1 - \delta(u - \hat{u}))(v^2 + \hat{v}^2) \end{aligned} \tag{20}$$

The mixed discrete/continuous nature of the source alphabet $\mathcal{X} = \mathcal{U} \times \mathcal{V} = \{1, 2, \ldots, N\} \times \mathbb{R}$ makes it difficult to compute (or make an educated guess at) the rate distortion function. Therefore we approach the problem by constructing several asymptotically achievable upper bounds. Asymptotic in this context means that we consider random codes for blocks of n outputs (u_i, v_i) from the spike process, with the block length n going to infinity. Furthermore we will work on the inverse function, that is the distortion rate function $D(R)$. For a memoryless Gaussian process it is $D_G(R) = \sigma^2 2^{-2R}$, see for example [3]. The first bound is derived by distributing the available rate uniformly among the N positions, thus coding for a Gaussian R.V. in each position:

$$D(R) \leq D_G(R/N) = \sigma^2 2^{-2\frac{R}{N}}. \tag{21}$$

Since this does not consider position coding, we will call (21) the *linear approximation bound*. Another bound, which is tight at high rates, is derived by coding the exact position with $\log N$ bits, and then using the remaining rate, if any, to code the Gaussian R.V. V:

$$D(R) \leq \min\{\sigma^2, D_G(R - \log N)\} = \sigma^2 \min\{1, 2^{-2(R-\log N)}\}. \tag{22}$$

This second bound suggests a two-part coding scheme as follows: suppose the block length n is large enough, so that we can encode the positions of $k \leq n$ spikes with $\log N$ bits each, and split the remaining rate equally among the k Gaussian R.V.s. If we pick an a-priori fixed set of spikes, e.g. the first k, we have $nR = k(\log N + R_V)$, where R_V is the rate assigned to each V_i. For $n \to \infty$ we can achieve any (rational) proportion $\mu = \frac{k}{n}$ of coded spikes. Obviously $0 \leq \mu \leq 1$. The rate available for coding V is $R_V = \mu^{-1}R - \log N$ and must be non-negative, yielding the condition

$$0 \leq \mu \leq \min\{1, R/\log N\} \tag{23}$$

The average distortion of such a two-part code will be

$$\mathrm{E}\rho(U, V; \widehat{U}, \widehat{V}) = \mu D_G(R_V) + (1-\mu)\sigma^2 = (\mu N^2 2^{-2\mu^{-1}R} + 1 - \mu)\sigma^2. \tag{24}$$

To minimize the average distortion over μ we compute the zero of the partial derivative $\frac{\partial}{\partial\mu}\mathrm{E}\rho(\cdot) = \sigma^2[N^2 2^{-2\mu^{-1}R}(1 + 2\ln(2)\mu^{-1}R) - 1]$ The solution $\mu = -\frac{2\ln(2)R}{W_{-1}(-e^{-1}N^{-2})+1}$ involves Lambert's W function, which satisfies $W(x)e^{W(x)} = x$. The subscript -1 indicates the second real branch of the W function, taking values on $[-1, -\infty[$. We notice that $\mu^{-1}R$ depends only on N. The condition (23) is met for all $N \geq 2$. If we plug this solution into (24), we get

$$D(R) \leq \min_{0\leq\mu\leq 1} \mathrm{E}\rho(\cdot) = \left(1 + \frac{2\ln(2)}{W(-e^{-1}N^{-2})}R\right)\sigma^2, \tag{25}$$

which is an upper bound to $D(R)$ as long as $\mu \leq 1$. If the rate is increased beyond that point, μ remains equal to 1, and we have $\mathrm{E}\rho(\cdot) = D_G(R - \log N)$ as in (22).

In hindsight we have obviously found the time sharing (a.k.a. multiplexing) solution between $(R, D) = (0, \sigma^2)$ and the point on (22) whose tangent goes through $(0, \sigma^2)$, as can be easily verified from the solution.

An obvious extension of this "first k" bound is to code those k spikes that are largest in magnitude. With an additional $nh(\mu) + 1$ bits we specify which k out of n spikes are coded (we will neglect the +1 term). The rate available to code each spike value is now $R_V = \mu^{-1}[R - h(\mu)] - \log N$. But the average distortion cannot be computed as in (24), since the $k = \mu n$ largest Gaussian spike values will have a variance larger than σ^2 for $0 < \mu < 1$. We have used a computer simulation to approximate the functions $v_1(\mu)$ and $v_2(\mu)$ which give the per-spike variances of the

$k = \mu n$ largest and $n-k = (1-\mu)n$ smallest spikes, respectively (in the limit of large block sizes, $\mu v_1(\mu) + (1-\mu)v_2(\mu) = \sigma^2$ holds). The average distortion now becomes (compare with (24)):

$$\mathrm{E}\rho(U,V;\widehat{U},\widehat{V}) = \mu v_1(\mu) N^2 2^{-2\mu^{-1}[R-h(\mu)]} + (1-\mu)v_2(\mu). \tag{26}$$

We get the "k largest" upper bound on $D(R)$ by sweeping μ from 0 to 1 and computing the convex hull of the curves (26) in the (R,D) plane.

Another way to look at this bound is by considering the expected number of coefficients with magnitudes above a threshold T. In the limit μ and $v_{1,2}$ become moments of a Gaussian R.V. restricted to the outside of $[-T,T]$ (inside for v_2). We were able to find a parametric expression for the achievable (R,D) points, but unfortunately the formulas are too large to fit in these proceedings. There is very good agreement between the analytic bound and the above numerical simulation approach.

As can be seen from Figure 2, the "k largest" bound is better than all other bounds. At high rates, all bounds except linear approximation (21) coincide with (22), as expected. For small N and low rates, linear approximation beats the "first k" approach by a small margin.

It is fairly easy to extend these upper bounds to multiple Gaussian spikes, since we are still using lossless position coding to construct them. That is, only the expression for R_V will change.

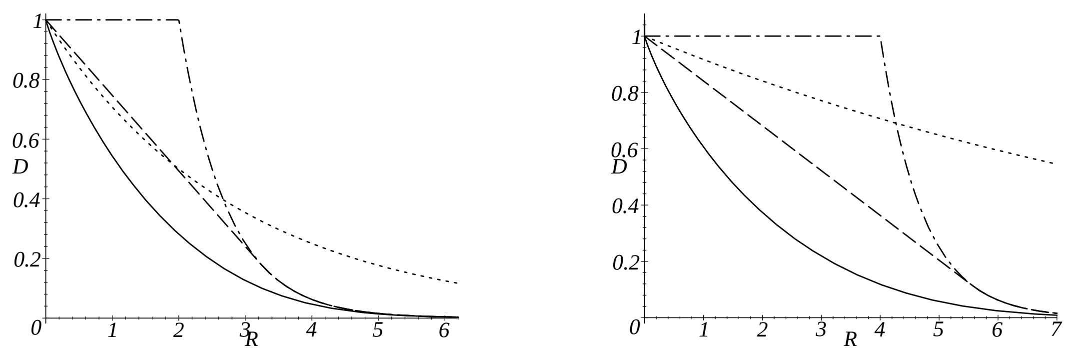

Figure 2: Upper bounds on $D(R)$ of a Gaussian spike with $\sigma = 1$ and $N = 4$ (at left), resp. $N = 16$ (at right). Dotted: linear approximation (21), dash-dotted: (22), dashed: "k first" (25), solid: "k largest" bound

4 Conclusion

The analysis of the Hamming spike case was mainly motivated by its relative simplicity. But it could be applied to *lossy* coding of the significance bitmaps of (embedded) image coding schemes such as SPIHT [8]. The main problem is to account for the distortion in the embedded scalar quantization which runs in parallel with the significance bitmap encoding. Another application of Hamming spikes is related to universal lossy source coding. To code a block of N symbols from a memoryless binary source with unknown distribution p, we count the number of ones (i.e. K) and encode it with $\log N$ bits. After that we use a random code which achieves $D(R - \log N)$ for the given N and K. From there we can get pointwise bounds similar to those presented in [5].

The Gaussian spike results obtained so far show the same kind of bi-modal rate distortion behavior as in Mallat and Falzon's analysis [6]. The main difference is that in their paper, $D(R)$ is found to decay approximately as R^{-1} at low rates, while we aren't yet able to give such a simple characterization of the low rate decay. But Figure 2 shows clearly that the "k largest" method, which is common to many practical compression schemes, pays off at low rates.

Our ongoing work is aimed at extending these results to generalized Gaussian processes and at incorporating the dependencies across scales of a wavelet transform. And of course the true rate distortion function of Gaussian spikes is still to be found.

References

[1] T. Berger, "Rate Distortion Theory", Prentice-Hall, 1971.

[2] A. Cohen, I.Daubechies, O. Guleryuz and M. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies", 1997, preprint.

[3] T. Cover, J. Thomas, "Elements of Information Theory", Wiley, 1991.

[4] D.L. Donoho, M. Vetterli, R.A. DeVore and I. Daubechies, "Data Compression and Harmonic Analysis", IEEE Trans. Inform. Theory, vol. 44, pp. 2435-2476, October 1998.

[5] T. Linder, G. Lugosi, K. Zeger, "Fixed-Rate Universal Lossy Source Coding and Rates of Convergence for Memoryless Sources", IEEE Trans. Inform. Theory, vol. 41, pp. 665-675, May 1995.

[6] S. Mallat and F. Falzon, "Analysis of Low Bit Rate Image Transform Coding", IEEE Trans. Signal Processing, vol. 46, pp. 19827-1042, April 1998.

[7] S. Mallat, "A Wavelet Tour of Signal Processing", Academic Press, 1998.

[8] A. Said and W.A. Pearlman, "A new fast and efficient image codec on set partitioning in hierarchical trees", IEEE Trans. on CSVT, Vol. 6, pp. 243-250, June 1996.