Analysis of Low Bit Rate Image Transform Coding

Stéphane Mallat [1] and Frédéric Falzon

To appear in IEEE Transactions on Signal Processing

April 1998

Abstract

Calculations based on high resolution quantizations prove that the distortion rate $D(\bar{R})$ of an image transform coding is proportional to $2^{-2\bar{R}}$ when $\bar{R}$ is large enough. In wavelet and block cosine bases, we show that if $\bar{R} < 1$ bit/pixel then $D(\bar{R})$ varies like $\bar{R}^{1-2\gamma}$, where γ remains of the order of 1 for most natural images. The improved performance of embedded codings in wavelet bases is analyzed. At low bit rates, we show that the compression performance of an orthonormal basis depends mostly on its ability to approximate images with a few non-zero vectors.

1 Introduction

If the signals to be encoded are realizations of a Gaussian process, under the high resolution quantization hypothesis, we know nearly everything about the performance of a transform coding. For an average of $\bar{R}$ bits per pixel, the mean-square error $D(\bar{R})$ varies proportionally to $2^{-2\bar{R}}$ with a constant that depends on the bit allocation and the basis. Current image transform coders operate below 1 bit per pixel. For such low bit rates, the high resolution quantization assumption yields an incorrect estimate of the distortion rate $D(\bar{R})$. In this range, we show that $D(\bar{R})$ depends mostly on the error D_0 when approximating signals with a limited number of vectors selected from the orthogonal basis. We verify that in wavelet and block cosine bases $D(\bar{R})$ varies like $\bar{R}^{1-2\gamma}$, where γ remains of the order of 1 for most "natural" images.

Transform coding algorithms can be improved by an embedding strategy which sends the larger amplitude coefficients first and then progressively refines their quantization. This improvement is analyzed mathematically and evaluated numerically for the wavelet zero-tree algorithm of Shapiro

[1]This work was supported by the French Centre National d'Etudes Spatiales and the AFOSR grant F49620-96-1-0455.

Stéphane Mallat is with Ecole Polytechnique, 91128 Palaiseau, France and with the Courant Institute, New York University, NY 10012.

Frédéric Falzon was at the INRIA, Sophia Antipolis, France, and is now at Alcatel Alsthom Recherche, 91460 Marcoussis, France.

[13]. Embedded coders outperform classical transform coders when there is some prior information on which basis vectors produce large, average or small decomposition coefficients for typical signals.

This article begins with a brief review of entropy constrained scalar quantizers and high bit rate transform coding. Section 3 analyzes the distortion rate at low bit rates and gives numerical examples with a wavelet transform coder and JPEG. The performance of embedded transform coders and their applications to wavelets is studied in Section 4.

2 High Bit Rate Compression

The class of signals to be encoded is represented by a random vector Y of size N. Although these signals may be multidimensional like images, they are indexed by an integer n to simplify notations: $Y[n]$. A transform coder decomposes these signals in an orthonormal basis $\mathcal{B} = \{g_m\}_{0 \le m < N}$

$$Y = \sum_{m=0}^{N-1} A[m]\, g_m .$$

Each coefficient $A[m]$ is a random variable defined by

$$A[m] = \langle Y, g_m \rangle = \sum_{n=0}^{N-1} Y[n]\, g_m^*[n].$$

To construct a finite code, each coefficient $A[m]$ is approximated by a quantized variable $\hat{A}[m]$. We concentrate on scalar quantizations which are most often used for transform coding. The next section reviews important results concerned with minimizing the quantization error.

2.1 Entropy Constrained Scalar Quantization

A scalar quantizer Q approximates a real random variable X by a quantized variable $\hat{X} = Q(X)$, which takes its values in a finite set. Suppose that X takes its values in $[a, b]$, which may correspond to the whole real axis. We decompose $[a, b]$ into K intervals $(y_{k-1}, y_k]_{1 \le k \le K}$ of variable lengths, with $y_0 = a$ and $y_K = b$. If $x \in (y_{k-1}, y_k]$ then $Q(x) = x_k$. We denote

$$p_k = \Pr\{X \in (y_{k-1}, y_k]\} = \Pr\{\hat{X} = x_k\}.$$

The Shannon theorem [5] proves that the entropy

$$\mathcal{H}(\hat{X}) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

is a lower bound of the average number of bits per symbol used to encode the values of $\hat{X}$. Arithmetic entropy coding [16] achieve an average bit rate that can be arbitrarily close to the entropy lower bound, so we shall consider that this lower bound is reached. An *entropy constrained scalar quantizer* is designed to minimize $\mathcal{H}(\hat{X})$ for a fixed mean-square distortion $D = \mathsf{E}\{(X - \hat{X})^2\}$.

Let $p(x)$ be the probability density of the random source X. The differential entropy of X is defined by

$$\mathcal{H}_d(X) = -\int_{-\infty}^{+\infty} p(x) \log_2 p(x)\, dx.$$

A quantizer is said to have a *high resolution* if $p(x)$ is approximately constant on each quantization bin $(y_{k-1}, y_k]$ of size $\Delta_k = y_k - y_{k-1}$. This is the case if the sizes Δ_k are sufficiently small relative to the rate of variation of $p(x)$. The following theorem [5] proves that uniform quantizers are optimal among high resolution quantizers. It is equivalent to the result of Girsh and Pierce [6], which proves the asymptotic optimality of uniform quantizers when the sizes of the quantization bins tend to zero.

Theorem 1 *If Q is a high resolution quantizer with respect to $p(x)$ then*

$$\mathcal{H}(\hat{X}) \geq \mathcal{H}_d(X) - \frac{1}{2}\log_2(12D). \tag{1}$$

This inequality is an equality if and only if Q is a uniform quantizer, in which case $D = \frac{\Delta^2}{12}$.

For a fixed distortion D, under the high resolution quantization hypothesis the minimum average bit rate $R_X = \mathcal{H}(\hat{X})$ is therefore achieved by a uniform quantizer and

$$R_X = \mathcal{H}_d(X) - \log_2 \Delta. \tag{2}$$

The distortion rate is obtained by taking the inverse of (1):

$$D(R_X) = \frac{1}{12}\, 2^{2\mathcal{H}_d(X)}\, 2^{-2R_X}. \tag{3}$$

Farvardin and Modestino [4] proved that even though the high resolution quantization hypothesis may not hold, for a large class of probability distribution including generalized Gaussians, the uniform quantizer yields a distortion rate that is close to the optimal quantizer if the number of quantization bins K is large enough.

2.2 Distortion Rate

Let us optimize the transform coding of $Y = \sum_{m=0}^{N-1} A[m]\, g_m$. The average bit budget to encode $\hat{A}[m] = Q(A[m])$ is $R_m = \mathcal{H}(\hat{A}[m])$. For a high resolution quantization, Theorem 1 proves that the error $D_m = \mathsf{E}\{|A[m] - \hat{A}[m]|^2\}$ is minimized when using a uniform scalar quantization. An optimal bit allocation minimizes the total number of bits $R = \sum_{m=0}^{N-1} R_m$ for a specified total error

$D = \sum_{m=0}^{N-1} D_m$. Let $\bar{R} = \frac{R}{N}$ be the average number of bit per sample. With Lagrange multipliers we verify that $\bar{R}$ is minimum if all D_m are equal [7], in which case

$$D(\bar{R}) = \frac{N}{12} 2^{2\overline{\mathcal{H}}_d} 2^{-2\bar{R}}, \tag{4}$$

where $\overline{\mathcal{H}}_d$ is the averaged differential entropy

$$\overline{\mathcal{H}}_d = \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{H}_d(A[m]).$$

A mean-square error D is not always a good measurement of visual degradations in images. In particular, we are often less sensitive to high frequency distortions as opposed to lower frequencies. A weighted mean-square norm takes into account this sensitivity by emphasizing low-frequency errors. Suppose that g_m is a vector whose Fourier transform is localized in a frequency neighborhood that depends on m, as in a block cosine basis or in a wavelet basis. A weight w_m is adjusted depending upon our visual sensitivity in this frequency range. The resulting weighted distortion is

$$D^w = \sum_{m=0}^{N-1} w_m^2 \, D_m = \sum_{m=0}^{N-1} D_m^w$$

where $D_m^w = w_m^2 \, D_m$ is the mean-square quantization error of $A^w[m] = w_m \, A[m]$. The previous bit allocation result applied to the coefficients $A^w[m]$ proves that D^w is minimized by quantizing uniformly $A^w[m]$ with a bin size Δ, which is equivalent to quantizing $A[m]$ with a bin size $\Delta_m = \frac{\Delta}{w_m}$. In the rest of the paper, we choose $w_m = 1$ and thus evaluate the error with a standard mean-square norm as opposed to a weighted norm. All calculations are easily extended to any other choice of weights by replacing $A[m]$ by $w_m \, A[m]$.

The distortion rate (4) depends on the orthonormal basis $\mathcal{B}$ through $\overline{\mathcal{H}}_d$. In general, it is difficult to find $\mathcal{B}$ which minimizes $\overline{\mathcal{H}}_d$ because the probability density of $A[m] = \langle Y, g_m \rangle$ may depend on g_m in a complicated way. If Y is a Gaussian random vector then the coefficients $A[m]$ are Gaussian random variables in any basis. In this case, the probability density of $A[m]$ depends only on the variance σ_m^2 and one can verify that

$$\mathcal{H}_d\Big(A[m]\Big) = \log_2 \sigma_m + \log_2 \sqrt{2\pi \mathrm{e}}.$$

Inserting this expression in (4) yields

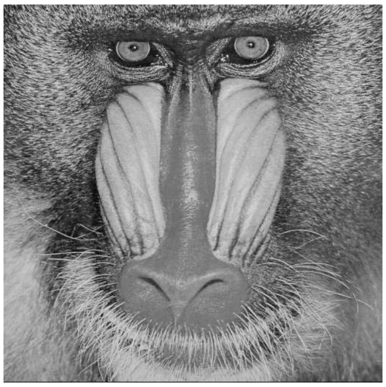
$$D(\bar{R}) = N \, \frac{\pi \, \mathrm{e}}{6} \left(\prod_{m=0}^{N-1} \sigma_m^2 \right)^{1/N} 2^{-2\bar{R}}. \tag{5}$$

One can prove that $\prod_{m=0}^{N-1} \sigma_m^2$ is minimum if and only if $\mathcal{B}$ is a Karhunen-Loève basis of Y [5], which means that $\mathcal{B}$ diagonalizes the covariance matrix of Y. The transform coding of a Gaussian process is thus optimized in a Karhunen-Loève basis. When Y is not Gaussian, the Karhunen-Loève basis is a priori not optimal anymore. This is the case for images which cannot be considered as realizations of a Gaussian process.

Let us describe a simple wavelet transform coder for images. Separable wavelet bases of images include three wavelets with horizontal, vertical or diagonal orientations [10], indexed by $1 \leq k \leq 3$. At an orientation k and scale 2^j, the wavelet vector $g_m = \psi^k_{j,p,q}$ is approximately centered at $(2^j p, 2^j q)$, with a square support whose size is proportional to 2^j. At high bit rates, we saw that the distortion rate is optimized by quantizing uniformly all decomposition coefficients. The domains where the image has smooth grey level variations yield small amplitude wavelet coefficients that are quantized to zero. To improve the efficiency of this transform coding, the wavelet coefficients are scanned in a predefined order and the position of zero versus non-zero quantized coefficients is recorded with a run-length coding that is entropy encoded. In the same scanning order, the amplitude of the non-zero quantized coefficients are also entropy encoded with a Huffman or an Arithmetic coding.

Lena

Barbara



Mandrill

Aerial

Figure 1: Four test images for numerical experiments.

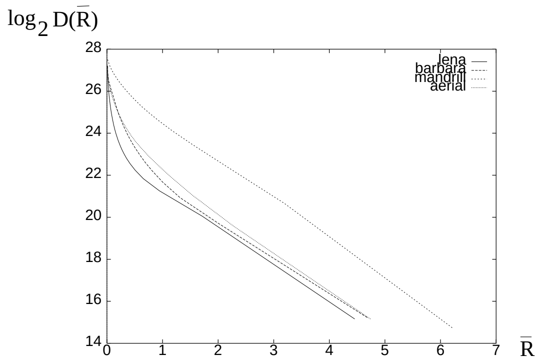


Figure 2: Log distortion rate curve for the wavelet transform coding of each test image.

Figure 2 gives the distortion $\log_2 D(\bar{R})$ of this wavelet transform coding for the test images shown in Figure 1. These numerical experiments are performed with an orthogonal cubic spline Battle-Lemarié wavelet [10]. The distortion rate formula (4) predicts that

$$\log_2 D(\bar{R}) = 2\overline{\mathcal{H}_d} + \log_2(\frac{N}{12}) - 2\bar{R} \ ,$$

where $\overline{\mathcal{H}_d}$ is the average differential entropy of the wavelet coefficients at all scales and positions. This formula implies that $\log_2 D(\bar{R})$ should decay with a slope of -2 as a function of $\bar{R}$. This is indeed verified in Figure 2 for $\bar{R} \geq 1$, but not for $\bar{R} < 1$ where $\log_2 D(\bar{R})$ has a much faster decay. At low-bit rates $\bar{R} < 1$ the distortion rate formula (4) is not valid because the high resolution quantization assumption does not hold. Wavelet transform codings are most often used for $\bar{R} < 1$ because they recover images of nearly perfect visual quality up to $\bar{R} = 0.5$ bits per pixel. The next section studies the distortion rate at these low bit rates.

3 Low Bit Rate Compression

At low bit rates, the decomposition coefficients of an image in an orthonormal basis are coarsely quantized. Since many coefficients are set to zero, the positions of zero versus non-zero quantized coefficients are stored in a binary significance map, which is recorded with a run-length coding or a more sophisticated zero-tree algorithm. The distortion rate theory previously described does not apply for two reasons. First, the high resolution quantization hypothesis does not hold because the quantization bins are large. Second, one cannot treat the total bit budget R as a sum of bits R_m allocated independently to each decomposition coefficient. Indeed, the encoding of the zero quantized coefficients through a significance map is a form of vector quantization, which relates the encoding of different coefficients.

To evaluate the distortion rate we cannot rely on a precise stochastic model for images. There is as yet no model that incorporates the full diversity of image structures, such as non-stationary textures and edges. To avoid this difficulty, we shall consider the signals as deterministic vectors whose decomposition coefficients in the basis $\mathcal{B}$ have a parameterized decay. The distortion rate is therefore not calculated with an ensemble average but for each signal f. The key result shows that this distortion rate depends mostly on the ability to precisely approximate f with a small number of vectors selected from $\mathcal{B}$. Low-bit rate image compressions in wavelet bases and block cosine bases illustrate the distortion rate results.

3.1 Distortion Rate

Let f be a signal decomposed in an orthonormal basis $\mathcal{B} = \{g_m\}_{0 \le m < N}$:

$$f = \sum_{m=0}^{N-1} a[m]\, g_m \quad \text{with } a[m] = \langle f, g_m \rangle.$$

The transform coder quantizes all coefficients and reconstructs

$$\hat{f} = \sum_{m=0}^{N-1} Q(a[m])\, g_m .$$

The coding error is

$$D = \|f - \hat{f}\|^2 = \sum_{m=0}^{N-1} |a[m] - Q(a[m])|^2. \tag{6}$$

We denote by $h[x]$ the discrete histogram of the N coefficients $a[m]$, normalized so that $\sum_x h[x] = 1$. The values of this histogram are interpolated to define a function $p(x) \ge 0$ for all $x \in \mathbf{R}$ such that $\int_{-\infty}^{+\infty} p(x)\, dx = 1$. This $p(x)$ is the probability density of a random variable X. We suppose N to be sufficiently large and the histogram sufficiently regular so that for all functions $\phi(x)$ that appear in our calculations one has :

$$\frac{1}{N} \sum_{m=0}^{N-1} \phi(a[m]) = \sum_x \phi(x)\, h[x] \approx \int_{-\infty}^{+\infty} \phi(x)\, p(x)\, dx = \mathsf{E}\{\phi(X)\} . \tag{7}$$

This hypothesis holds for the histograms of the test images shown in Figure 3, as well as for most "natural" images. It is equivalent to the coefficients $a[m]$ being successive values of the random variable X. Applied to $\phi(x) = |x - Q(x)|^2$, (7) yields

$$\frac{D}{N} = \frac{1}{N} \sum_{m=0}^{N-1} |a[m] - Q(a[m])|^2 = \mathsf{E}\{|X - Q(X)|^2\}.$$

Let $\bar{R}$ be the average number of bits per coefficient to encode the $Q(a[m])$. If Q is a high resolution uniform quantizer with step size Δ then (3) implies a distortion rate formula similar to (4):

$$\frac{D(\bar{R})}{N} = \frac{1}{12}\, 2^{2\mathcal{H}_d(X)}\, 2^{-2\bar{R}} .$$

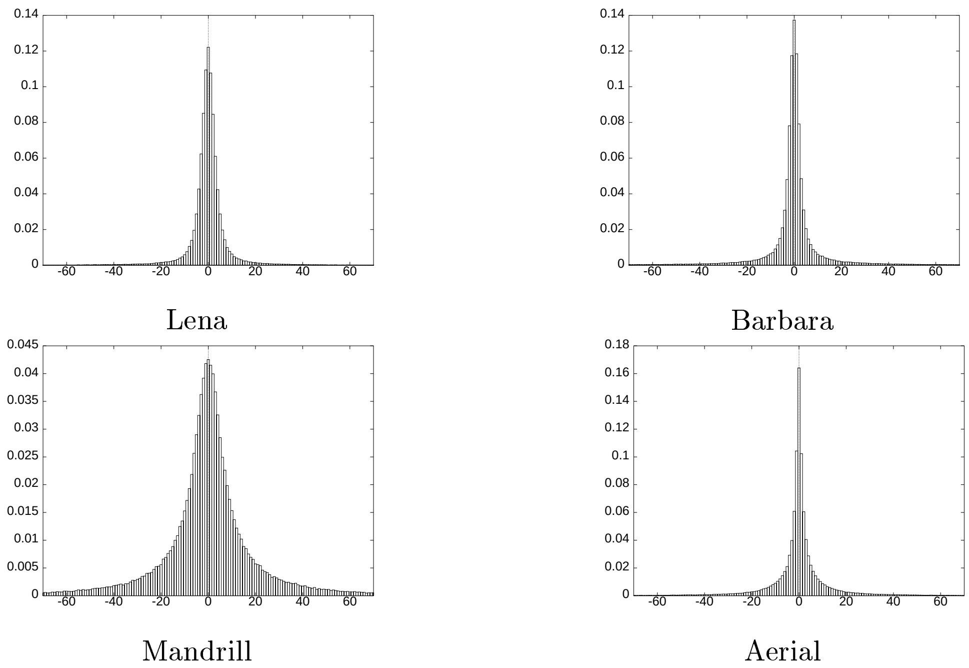


Figure 3: Normalized histograms of the cubic-spline wavelet coefficients of the test images.

If the basis $\mathcal{B}$ is chosen so that many coefficients $a[m] = \langle f, g_m \rangle$ are close to zero and few have a large amplitude then $p(x)$ has a sharp high peak at $x = 0$. This is the case for the histograms of wavelet coefficients shown in Figure 3. If Δ is large then $p(x)$ has important variations in the zero bin $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ where coefficients are quantized to zero. Hence the high resolution quantization hypothesis does not apply in this zero bin. This explains why $\log_2 D(\bar{R})$ shown in Figure 2 decays like $-2\bar{R}$ only for $\bar{R} \geq 1$.

If $p(x)$ is a laplacian distribution $p(x) = \frac{1}{\sigma\sqrt{2}} \mathrm{e}^{\frac{-\sqrt{2}|x|}{\sigma}}$ then Sullivan [14] proved that the optimal entropy constrained scalar quantizer in a nearly uniform quantizer. All the non-zero quantization bins have the same size Δ but the zero bin $[-T, T]$ is larger, with a ratio $\theta = \frac{T}{\Delta}$ that can be calculated. Sullivan's result does not apply to the quantization of wavelet image coefficients because Section 3.2 shows that $p(x)$ has a slower decay which is rational instead of exponential. Yet, efficient low bit rate wavelet transform coders are often implemented with a nearly uniform quantizer, whose zero bin $[-T, T]$ is larger than the other bins. This can be justified with a high resolution assumption outside the zero bin. For $|x| > T$, we shall consider that the relative variations of $p(x)$ in each quantization bin is sufficiently small to apply the high resolution hypothesis. This assumption is only an approximation but it captures sufficiently precisely the quantizer properties to obtain accurate calculations up to very low bit rates. For $x \in [-T, T]$, the high-resolution quantization hypothesis does not hold because $p(x)$ varies too much. Theorem 1 implies that outside $[-T, T]$

the optimal entropy constrained quantizer Q has bins of constant size Δ. The ratio $\theta = \frac{T}{\Delta}$ is a parameter that must be adjusted to minimize the overall distortion D.

Any coefficient $|a[m]| > T$ that is not quantized to zero is called a *significant coefficient*. Coding the position of non-zero quantized coefficients is equivalent to storing a binary *significance map* defined by

$$b[m] = \begin{cases} 0 & \text{if } |a[m]| \le T \\ 1 & \text{if } |a[m]| > T \end{cases} . \tag{8}$$

JPEG and the wavelet image coder of Section 2.2 use a run-length coding to store this significance map. More efficient zero-tree encoding techniques may also be used for wavelet significance maps [9].

Let R_0 be the total number of bits required to encode the significance map. Let M be the number of significant coefficients. There is a proportion $p = \frac{M}{N}$ of indices m such that $b[m] = 1$. An upper bound for R_0 is computed by supposing that there is no redundancy in the position of the 0 and the 1 in the significance map. The average number of bits to encode the position of one coefficient is then the entropy of a binary source with a probability $p = \frac{M}{N}$ to be equal to 1 and $1-p$ to be equal to 0:

$$\frac{R_0}{N} \le -p \log_2 p - (1-p)\log_2(1-p).$$

For $x \in (0,1]$ then $-x \log_2 x \le (1-x)\log_2 \mathrm{e}$ so the average number of bits per significant coefficient to encode the significance map is

$$r_0 = \frac{R_0}{M} \le \log_2 \frac{N}{M} + \log_2 \mathrm{e}. \tag{9}$$

For wavelet coefficients, when the proportion of significant coefficients $\frac{M}{N}$ is small, a run-length coding yields an average bit rate $r_0 = \frac{R_0}{M}$ which is much smaller than the upper bound (9), because of the redundancy in the positions of the zero coefficients. For large classes of images, numerical calculations show that $r_0 = \frac{R_0}{M}$ varies slowly relative to $\frac{M}{N}$.

The amplitude of the M significant coefficients is uniformly quantized with a step Δ and these quantized values are entropy encoded. Let us compute the total number of bits R_1 of the resulting entropy coding. For $M \gg 1$, the M significant coefficients $a[m]$ above T have a normalized histogram that is interpolated by

$$p_T(x) = \frac{N}{M}\, p(x)\, \mathbf{1}_{\{|x|>T\}} .$$

Let X_T be the random variable whose probability density is $p_T(x)$. Since the high resolution quantization hypothesis applies to significant coefficients, the average number of bits to encode the amplitude of each quantized significant coefficient, denoted r_1, is calculated from (2):

$$r_1 = \frac{R_1}{M} = \mathcal{H}_d(X_T) - \log_2 \Delta. \tag{10}$$

Overall, the transform coding requires $R = R_0 + R_1$ bits.

To estimate the quantization error, $D = \|f - \hat{f}\|^2$ in (6), insignificant coefficients quantized to zero are separated from significant coefficients: $D = D_0 + D_1$, where

$$D_0 = \sum_{|a[m]| \leq T} |a[m]|^2 \tag{11}$$

is the error due to quantizing insignificant coefficients, and

$$D_1 = \sum_{|a[m]| > T} |a[m] - Q(a[m])|^2 \tag{12}$$

is the error due to quantizing significant coefficients. The average quantization error $\frac{D_1}{M}$ per significant coefficient is calculated with the high resolution quantization assumption

$$\frac{D_1}{M} = \frac{1}{M} \sum_{|a[m]| > T} |a[m] - Q(a[m])|^2 = \mathsf{E}\{|X_T - Q(X_T)|^2\} = \frac{\Delta^2}{12}. \tag{13}$$

To compute the error due to quantizing insignificant coefficients we denote by f_M the approximation of f using the M vectors g_m of $\mathcal{B}$ such that $|a[m]| = |\langle f, g_m \rangle| > T$

$$f_M = \sum_{|a[m]| > T} a[m]\, g_m.$$

The signal f_M can also be interpreted as an approximation of f from the M vectors of $\mathcal{B}$ whose inner products with f have the largest amplitude. The distortion D_0 can be rewritten

$$D_0\Big(\frac{M}{N}\Big) = \|f - f_M\|^2 = \sum_{|a[m]| \leq T} |a[m]|^2 \tag{14}$$

In approximation theory, D_0 is called a *non-linear approximation error* because the M vectors are selected depending upon f as opposed to linear algorithms that approximate all signals with the same M vectors. Clearly D_0 decays when M increases, but how fast? This issue is a central question that is studied in approximation theory in relation to particular functional spaces [3, 11].

Let us sort the inner products $a[m]$ by their amplitude. The amplitude of the k^{th} coefficient is written

$$x\left(\frac{k}{N}\right) = |a[m_k]| \leq x\left(\frac{k+1}{N}\right) = |a[m_{k+1}]| \quad \text{for } 1 \leq k < N. \tag{15}$$

The approximation error D_0 is the sum of the $N - M$ squared coefficients of smaller amplitude

$$D_0\left(\frac{M}{N}\right) = \sum_{k=M+1}^{N} \left| x\left(\frac{k}{N}\right) \right|^2. \tag{16}$$

The error $D_0(\frac{M}{N})$ has a fast decay when $\frac{M}{N}$ increases if $x(z)$ decreases quickly when z increases.

Observe that $p(x)$ is related to the inverse $z(x)$ of $x(z)$ by

$$z(x) = 1 - \int_{-x}^{x} p(u)\, du. \tag{17}$$

The probability density $p(x)$ is defined as a function of a continuous variable by interpolating the normalized histogram of the decomposition coefficients $a[m]$ of f. This also defines a function $x(z)$ for any $z \in [0,1]$, which interpolates the values $x(\frac{k}{N})$.

To estimate the decay of $D_0(z)$ when z increases, a standard approximation theory approach computes the rational decay of the sorted coefficients, and hence compares $x(z)$ with $z^{-\gamma}$ for some $\gamma > 0$. For functions $f \in [0,N]^2$ decomposed in a wavelet orthonormal basis, the exponent γ characterizes particular functional spaces called Besov spaces [3]. To suppose that $x(z) = C\, z^{-\gamma}$ would clearly be too restrictive to model interesting classes of signals. We shall rather suppose that this exponent is slowly varying and define

$$\gamma(z) = -\frac{d \log_2 x(z)}{d \log_2 z}. \tag{18}$$

Figure 4 plots $\log_2 x(z)$ as a function of $\log_2 z$ for the wavelet coefficients and the block cosine coefficients of the test images. Observe that in both cases the slope $\gamma(z)$ varies slowly for $z \leq 2^{-1}$. This behavior is further discussed in Section 3.2.

To compute the distortion rate we shall assume that the second order derivative is bounded by a small $\epsilon > 0$

$$\left| \frac{d^2 \log_2 x(z)}{(d \log_2 z)^2} \right| \leq \epsilon \quad \text{for } z \in (0, \tfrac{1}{2}]. \tag{19}$$

We also suppose that

$$\inf_{z \in [0,1]} \gamma(z) > 0, \tag{20}$$

$$\frac{d^2 \log_2 x(z)}{(d \log_2 z)^2} \leq 0 \quad \text{for } z \in (0,1), \tag{21}$$

and that $p(x)$ is symmetric

$$p(x) = p(-x)\ . \tag{22}$$

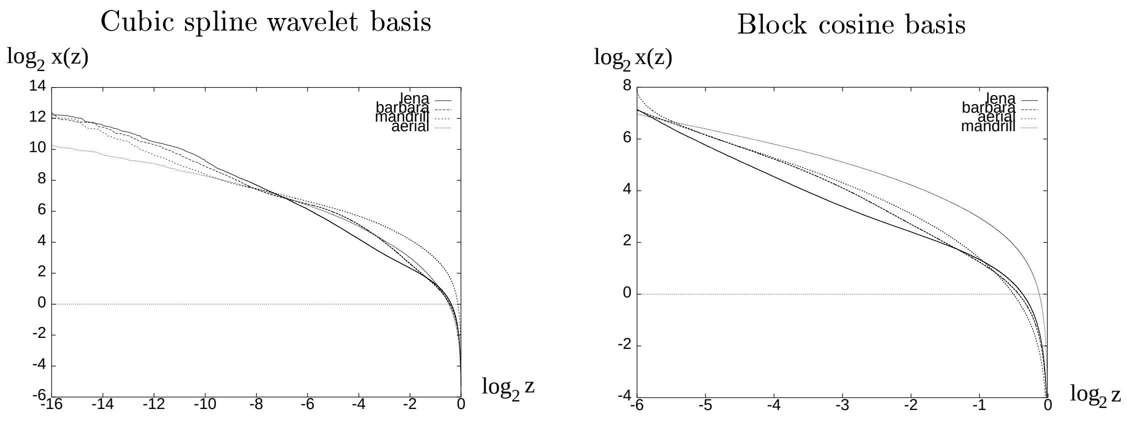
The four test images as well as most natural images have wavelet and block cosine coefficients that satisfy (19-22). The concavity (21) is a technical condition that is used to control corrective terms in distortion rate calculations, but which is generally satisfied. The symmetry (22) of the probability density is verified in Figure 3. Assuming (19-22), the following theorem relates the distortion rate $D(\bar{R})$ to the approximation error D_0, through parameters that are evaluated as a function of $\gamma(z)$ and the number M of significant coefficients.

Theorem 2 *Suppose that $x(z)$ satisfies (19-22). Let $\gamma_M = \gamma(\frac{M}{N}) > \frac{1}{2}$. If $\frac{M}{N} \leq \epsilon$ and $M \geq \frac{1}{\epsilon}$ then*

$$D(\bar{R}) = (1+K)\, D_0 \left(\frac{\bar{R}}{r_1 + r_0} \right) \tag{23}$$

with

$$K = \frac{D_1}{D_0} = \frac{2\gamma_M - 1}{12\,\theta^2} \left[1 + O\Big(\epsilon |\log_2 \epsilon|^2 + \epsilon^{2\gamma_M - 1} \Big) \right], \tag{24}$$



(a) (b)

Figure 4: Amplitude of the sorted decomposition coefficients of the test images in two bases.

and

$$r_1 = \frac{R_1}{M} = 1 + (1 + \gamma_M) \log_2 \mathrm{e} + \log_2 \gamma_M + \log_2 \theta + O(\epsilon). \tag{25}$$

Moreover, the derivative of $D_0(z)$ satisfies

$$\frac{d \log_2 D_0(\frac{M}{N})}{d \log_2 z} = (1 - 2\gamma_M) \left[1 + O\Big(\epsilon |\log_2 \epsilon|^2 + \epsilon^{2\gamma_M - 1}\Big)\right]. \tag{26}$$

The proof is in appendix A.1. To understand the implications of this theorem, these formulae are simplified with an approximation, and we neglect the corrective terms in ϵ. Since the second order derivative (19) remains small, the slope

$$\gamma_M = \frac{-\, d \log_2 x(\frac{M}{N})}{d \log_2 z} \tag{27}$$

varies slowly as a function of $\log_2 \frac{M}{N}$. In the compression range of interest, it will be considered constant: $\gamma_M \approx \gamma$. It follows from (24) and (25) that $K = \frac{D_1}{D_0}$ and $r_1 = \frac{R_1}{M}$ are also constant. We have already mentioned that this is also the case for $r_0 = \frac{R_0}{M}$. Hence $D(\bar{R})$ is calculated in (23) by scaling and multiplying the non-linear approximation error $D_0(z)$ by constant factors

$$D(\bar{R}) = (1 + K)\, D_0 \left(\frac{\bar{R}}{r_1 + r_0} \right). \tag{28}$$

Since $\gamma_M \approx \gamma$, (26) implies that $D_0(z) \sim z^{2\gamma - 1}$ so

$$D(\bar{R}) \sim \bar{R}^{1 - 2\gamma} .$$

This distortion rate decay is very different from the high resolution formula where $D(\bar{R}) \sim 2^{-2\bar{R}}$.

The distortion D in (28) depends essentially upon the approximation error D_0 of f from $M = \frac{\bar{R}}{r_1+r_0}$ vectors selected in the basis $\mathcal{B}$. To optimize the transform coding, the basis $\mathcal{B}$ must be able to approximate precisely each signal f with a small number of basis vectors. If we consider f as a realization of a random vector Y then we may wonder which is the basis which minimizes $\mathsf{E}\{D_0(\frac{M}{N})\}$ over all realizations. This problem is difficult because the M basis vectors are selected depending upon each realization f of Y. They are the ones which have the largest inner products with f. The energy compaction theorem [5] proves that the Karhunen-Loève basis is optimal for approximating Y from M vectors chosen once and for all, but it has no optimality property in this non-linear setting where the vectors depend upon each realization. In some cases, we know how to find bases that minimize the maximum error $D_0(\frac{M}{N})$ over a whole signal class. For example, wavelet bases are optimal in this min-max sense for piece-wise regular signals that belong to Besov spaces [3].

To optimize the quantization, the size of the zero bin $[-T, T]$ must be adjusted with respect to the other quantization bins of size Δ. To minimize the distortion D, we want to find $\theta = \frac{T}{\Delta}$ such that for a fixed $\bar{R}$

$$\frac{\partial D(\bar{R}, \theta)}{\partial \theta} = 0. \tag{29}$$

Appendix A.2 proves the following theorem that gives an analytic formula for θ.

Theorem 3 *Suppose that $x(z)$ satisfies (19-22) and that $r_0 = \frac{R_0}{M}$ is a constant independent of M. Let $\gamma_M = \gamma(\frac{M}{N}) > \frac{1}{2}$. If $\frac{M}{N} \leq \epsilon$ and $M \geq \frac{1}{\epsilon}$ then the optimal zero bin ratio is*

$$\theta = \sqrt{\frac{r_1 + r_0}{6 \ \log_2 \mathrm{e}} - \frac{2\,\gamma_M - 1}{12}} \left[1 + \mathcal{O}\left(\epsilon\right)\right]. \tag{30}$$

3.2 Wavelet Transform Coding

Wavelet bases are known to efficiently approximate piece-wise regular functions with a small number of non-zero wavelet coefficients [11]. Since images often include piecewise regular structures, wavelet bases are good candidates for building efficient image transform coders. The central assumption of Theorem 2 is that the sorted decomposition coefficients $x(z)$ of f in the basis $\mathcal{B}$ have a rational decay.

In wavelet bases, this is in accordance with asymptotic image models based on Besov spaces introduced for compression by DeVore, Jawerth and Lucier [3], and further studied in [1]. Let $f \in \mathbf{L^2}[0, N]^2$. If there exists $C > 0$ and $\gamma > \frac{1}{2}$ so that for all $k \geq 0$ the wavelet coefficient of f of rank k is bounded by $C\, k^{-\gamma}$ then f belongs to a family of Besov spaces whose indexes depend upon γ. Let us consider a piecewise regular image f, which is uniformly regular (Lipschitz $\alpha \geq 1$) inside the regions $\{\Omega_i\}_{1 \leq k \leq K}$ which partition $[0, N]^2$. This image has discontinuities along the borders of the Ω_i, which have a finite length. One can then prove [11] that the sorted wavelet coefficients decay like $C\,k^{-\gamma}$, with $\gamma = 1$. The discontinuities create large amplitude wavelet coefficient which are responsible for this decay exponent. This piecewise regular model applies to an image such as Lena because even the fur texture does not create enough large wavelet coefficients to modify the exponent $\gamma = 1$. On the other hand, the Mandrill image is composed of regions with highly irregular textures which create enough high amplitude wavelet coefficients to reduce the exponent γ. For finite images $k \leq N$, which is why we renormalize $z = \frac{k}{N}$ and compare the decay of $x(z)$ with $z^{-\gamma}$ when z increases in $[0, 1]$. Theorem 2 does not require that γ is a constant but (19) assumes that it varies slowly as a function of $\log_2 z$.

If $x(z) \sim z^{-\gamma}$ for x large enough then one can derive from (51) that $p(x) \sim x^{-1-\frac{1}{\gamma}}$. However, Figure 3 shows that $p(x)$ has an exponential decay when x is small. This can be explained by looking at the normalized histograms $p_j(x)$ of the wavelet coefficients $\langle f, \psi^k_{j,p,q} \rangle$ at a fixed scale 2^j, for all positions $0 \leq p, q \leq 2^{-j}N$ and orientations $1 \leq k \leq 3$. Such histograms are well modeled by generalized Gaussian distributions [10], which have an exponential decay and a variance which increases with the scale 2^j. It is the aggregation of these histograms that yield a global histogram $p(x)$ which has a rational decay for x sufficiently large [8]. However, the finite image resolution implies that wavelet coefficients are zero for $j < 0$. One can verify that the "border effect" created by the absence of finer scale wavelet coefficients implies that $p(x) \approx p_1(x)$ for x small. This explains the exponential decay of $p(x)$ when x is small, and the rapid variation of the slope γ for $\log_2 z \geq -1$ in Figure 4(a).

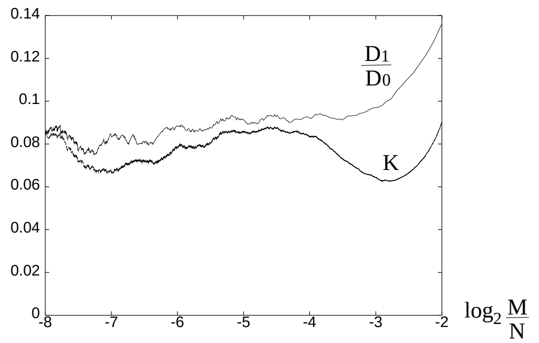


Figure 5: Comparison of $\frac{D_1}{D_0}$ with the theoretical estimate $K = \frac{2\,\gamma_M - 1}{12\,\theta^2}$ for the wavelet coding of Lena.

Let us now evaluate numerically the precision of the analytic formula given by Theorem 2. The wavelet transform coder is implemented with a quantizer whose zero bin $[-T, T]$ is twice as large as the other quantization bins, which means that $\theta = \frac{T}{\Delta} = 1$. This is a standard choice in most wavelet image transform coders. The significance map is stored with a run-length coding as explained in Section 2.2. The ratio $\frac{D_1}{D_0}$ was calculated numerically for the wavelet transform coding of Lena. Figure 5 compares this value with the theoretical estimate (24), where the corrective terms in ϵ are neglected: $K = \frac{2\,\gamma_M - 1}{12\,\theta^2}$. The slope γ_M in (27) is computed numerically from the sorted coefficients $x(z)$ of Lena shown in Figure 4(a). For $\frac{M}{N} \leq 2^{-4}$, Figure 5 shows that K closely approximates the true value of $\frac{D_1}{D_0}$. The error increases for $\frac{M}{N} > 2^{-4}$ because the Theorem hypothesis $\frac{M}{N} \leq \epsilon$ is not respected.

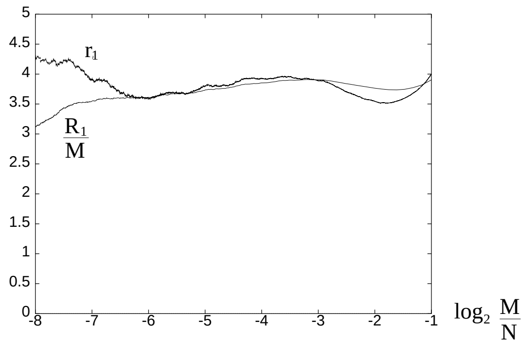


Figure 6: Comparison of $\frac{R_1}{M}$ with the theoretical estimate $r_1 = 1 + (1+\gamma_M)\log_2 \mathrm{e} + \log_2 \gamma_M + \log_2 \theta$ for the wavelet coding of Lena.

Figure 6 displays the value of $\frac{R_1}{M}$ computed numerically from the entropy of the quantized wavelet coefficients of Lena. The theoretical estimate (25)

$$r_1 = 1 + (1+\gamma_M)\log_2 \mathrm{e} + \log_2 \gamma_M + \log_2 \theta. \tag{31}$$

is plotted in the same figure. The curves remain close, which verifies the precision of this calculation.

To simplify the expression of the distortion rate, the slope γ_M is approximated by a constant $\gamma_M \approx 1$, which corresponds to piecewise regular image models. Although γ_M can differ from 1 in many images such as Mandrill, this approximation is justified by the small sensitivity of K and $r_1 + r_0$ with respect to fluctuations of γ_M around 1. Since $\theta = 1$, we get $K \approx \frac{1}{12}$ and $r_1 \approx 3.9$. Figure 7 displays

$$\frac{R}{M} = \frac{R_0}{M} + \frac{R_1}{M} = r_0 + r_1$$

which was computed numerically for the four test images. For $\frac{M}{N} \in [2^{-7}, 2^{-1}]$, the ratio $\frac{R}{M}$ can be approximated by a constant $r_0 + r_1 \approx 6.5$. The distortion D calculated in (23) is thus approximated by

$$\hat{D}(\bar{R}) = \left(1 + \frac{1}{12}\right) D_0 \left(\frac{\bar{R}}{6.5}\right). \tag{32}$$

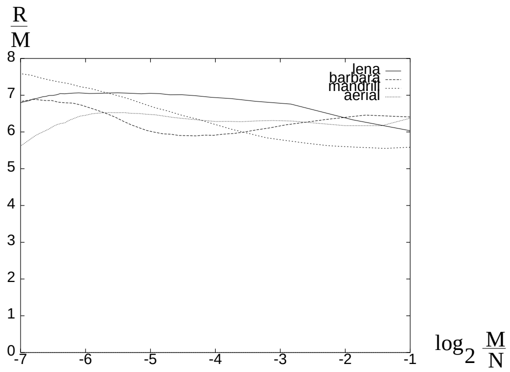


Figure 7: Variations of $\frac{R}{M}$ for the wavelet transform coding of the test images.

Figure 8 compares the Peak Signal to Noise Ratio

$$P_{SNR}(D) = 10\ \log_{10} \frac{N\ 255^2}{D},$$

which was calculated numerically for the four test images, with the approximated $P_{SNR}(\hat{D})$ derived from (32). Observe that $\hat{D}(\bar{R})$ gives a remarkably precise evaluation of the true distortion rate $D(\bar{R})$, despite the fact that γ_M is not exactly equal to 1.

The increment of $P_{SNR}(D)$ for each additional bit is calculated by inserting (26) in (30) while neglecting the variations of K and $r_1 + r_0$

$$\frac{dP_{SNR}(D)}{d \log_2 \bar{R}} \approx (2\gamma_M - 1)\ 10\ \log_{10} 2.$$

For Lena, $\gamma_M = 1$ so $P_{SNR}(D)$ increases by 3db for each additional bit, which is indeed verified in Figure 8. For the three other images, the variations of γ_M can not be neglected over the whole compression range $\bar{R} \in [2^{-6}, 1]$ and Figure 8 shows that $P_{SNR}(D)$ has a slope which varies slowly with $\log_2 \bar{R}$.

Theorem 3 gives an analytical expression which computes $\theta = \frac{T}{\Delta}$ that minimizes the distortion D. For $r_1 + r_0 = 6.5$ and $\gamma_M \approx 1$ we get $\theta = 0.81 \approx 1$. This theoretical estimate is precisely the choice that is most often used in wavelet compression softwares, after ad-hoc numerical trials.

To optimize the wavelet transform coder, the distortion rate (32) shows that one must choose a wavelet basis that gives precise approximations of images with few wavelet coefficients, in order to maintain a small approximation error $D_0(z)$. This essentially depends on the support size and the number of vanishing moments of the wavelets [11]. The optimization of the wavelet basis may depend upon the particular class of images to be encoded.

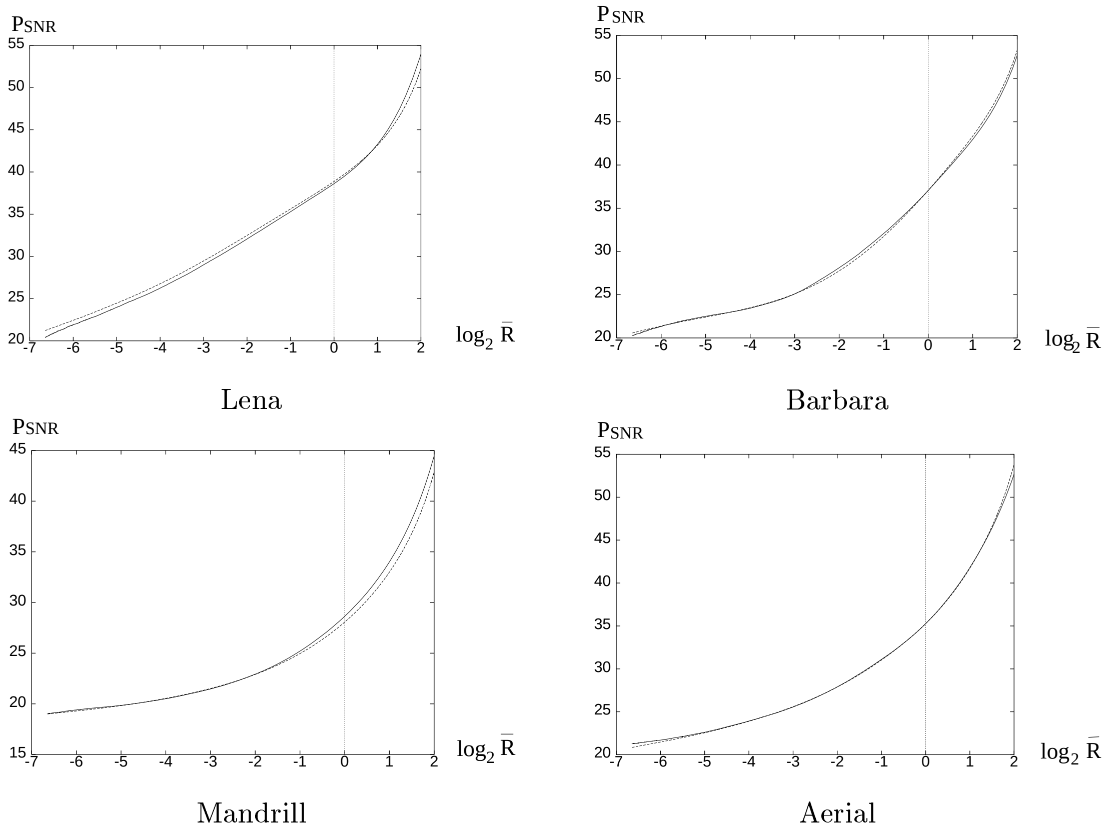


Figure 8: $P_{SNR}(D)$ (solid line) and $P_{SNR}(\hat{D})$ (dashed line) for a wavelet coder, with $\hat{D} = (1 + \frac{1}{12})\, D_0(\frac{\bar{R}}{6.5})$.

3.3 JPEG Image Coding

The JPEG image standard decomposes an image in a block cosine basis $\mathcal{B}$. Images of N^2 pixels are divided in $\frac{N^2}{64}$ blocks of $L = 8$ by 8 pixels. Figure 4(b) shows that the decay of the sorted local cosine coefficients satisfies the assumptions of Theorem 2 for the four test images.

JPEG uniformly quantizes the block cosine coefficients. In each block of 64 pixels there is one DC coefficient, which is proportional to the average image value over the block. Instead of directly quantizing this DC coefficient, JPEG quantizes the differences between the DC values of two adjacent blocks. Since the amplitudes of the DC coefficients are not directly quantized, their values are not included in the sorted coefficients shown in Figure 4(b). A significance map gives the position of zero versus non-zero quantized coefficients. It is encoded with a run-length coding, which scans the 64 cosine coefficients of each block. The following numerical calculations are done with a baseline JPEG compression coder. Following our explanation in Section 2.2, the weights w_m used to improve the visual quality of JPEG images are set to 1, which maintains a uniform quantization with same bin size Δ for all coefficients.

Since JPEG uniformly quantizes the decomposition coefficients, the zero bin $[-T, T]$ is equal to other bin sizes Δ and hence $\theta = \frac{T}{\Delta} = \frac{1}{2}$. For Lena, Figure 9 compares the ratio $\frac{D_1}{D_0}$ computed numerically and the theoretical estimate derived from Theorem 2

$$K = \frac{2\gamma_M - 1}{12\ \theta^2} = \frac{2\gamma_M - 1}{3}.$$

The slope γ_M is calculated from the coefficient decay shown in Figure 4(b). Both curves are very close up to $\frac{M}{N} \geq 2^{-3}$. For $\frac{M}{N} \geq 2^{-3}$, the estimation error of $\frac{D_1}{D_0}$ increases because it does not respect the theorem hypothesis that $\frac{M}{N} \leq \epsilon$.

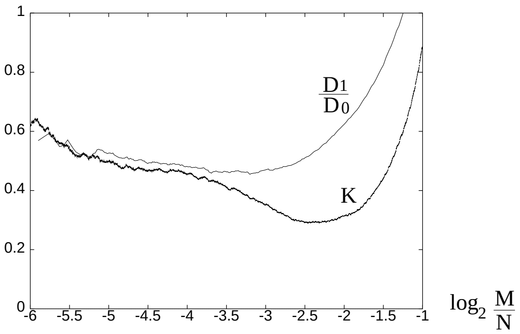


Figure 9: Comparison of $\frac{D_1}{D_0}$ with the theoretical estimate $K = \frac{2\,\gamma_M - 1}{12\,\theta^2}$, for the JPEG coding of Lena.

JPEG uses a mixed format that encodes the run-length coding for significant coefficients and the amplitude of these significant coefficients together [15]. It is therefore not possible to compute R_0 and R_1 separately. Figure 10 displays $\frac{R}{M} = r_0 + r_1$ as a function of $\log_2 \frac{M}{N}$ for the four test images. If $\frac{M}{N} \in [2^{-5}, 2^{-1}]$ then $r_0 + r_1 \approx 5.5$.

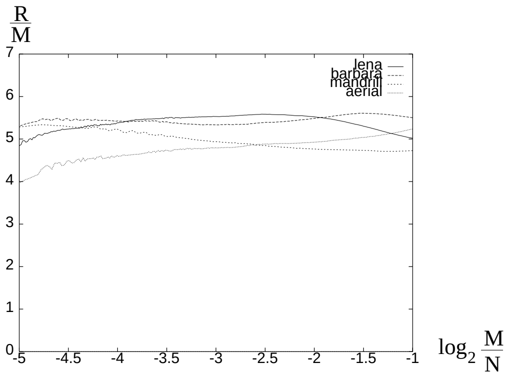


Figure 10: Variations of $\frac{R}{M}$ for the JPEG coding of the test images.

In the compression range of JPEG, the slope γ_M remains slightly above 1. Approximating $\gamma_M \approx \frac{5}{4}$ yields $K \approx \frac{1}{2}$. The distortion rate formula calculated with (23)

$$D(\bar{R}) = (1 + K)\, D_0 \left(\frac{\bar{R}}{r_0 + r_1} \right)$$

is thus approximated by

$$\hat{D}(\bar{R}) = \left(1 + \frac{1}{2}\right) D_0 \left(\frac{\bar{R}}{5.5} \right). \tag{33}$$

Figure 11 compares $P_{SNR}(D)$ calculated numerically with JPEG software and its approximation $P_{SNR}(\hat{D})$ derived from (33). It shows that the distortion rate formula (33) is a precise approximation over the whole compression range of JPEG, and that we can neglect the variations of K and $r_0 + r_1$.

4 Embedded Transform Coding

For rapid transmission or fast image browsing from a data base, one should provide a coarse signal approximation quickly, and then progressively enhance it as more bits are transmitted. Embedded coders offer this flexibility by grouping the bits in order of significance. The decomposition coefficients are sorted and the first bits of the largest coefficients are sent first. An image approximation can be reconstructed at any time from the bits already transmitted. Embedded coders can take advantage of any prior information on the location of large versus small coefficients. Such prior

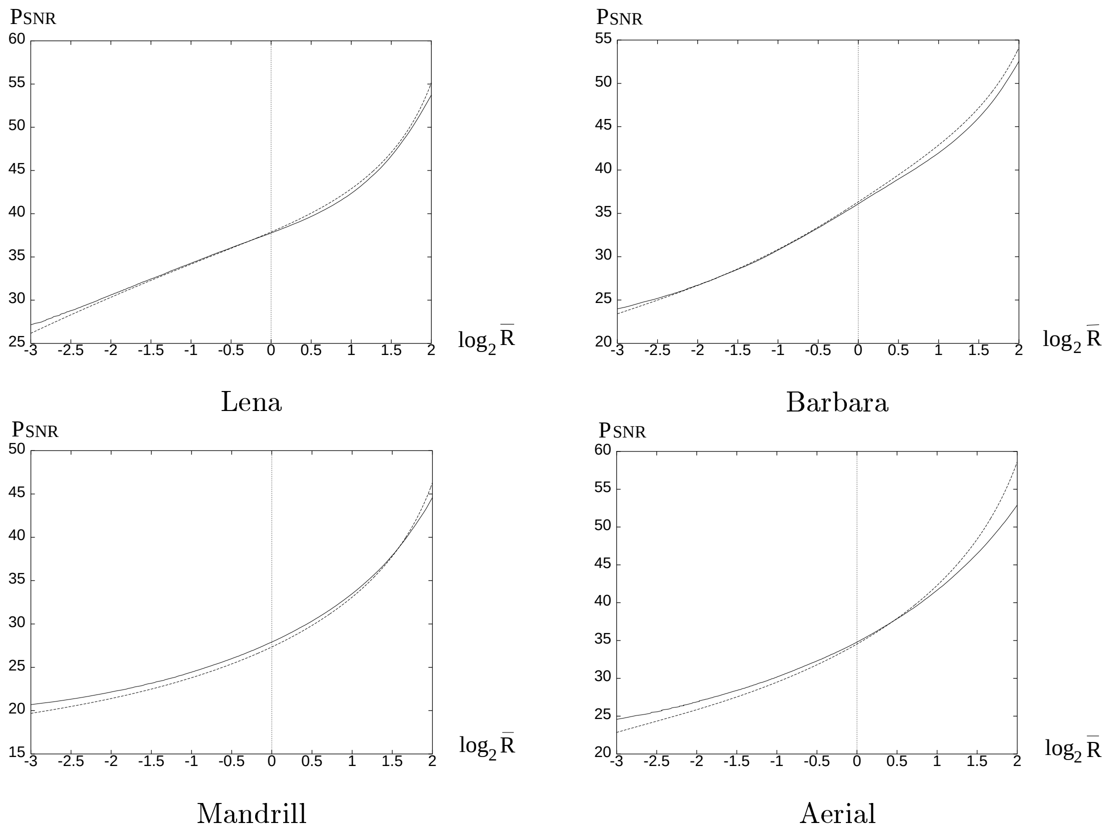


Figure 11: $P_{SNR}(D)$ (solid line) and $P_{SNR}(\hat{D})$ (dashed line) for JPEG, with $\hat{D} = (1 + \frac{1}{2})\, D_0(\frac{\bar{R}}{5.5})$.

information is available for natural images decomposed in wavelet bases. As a result, an implementation with zero-trees designed by Shapiro [13] yields better compression rates than classical wavelet transform coders.

The decomposition coefficients $a[m] = \langle f, g_m \rangle$ are partially ordered by grouping them in sets $\mathcal{S}_k$ of indices defined for any $k \in \mathbf{Z}$ by

$$\mathcal{S}_k = \{m \ : \ 2^k \le |a[m]| < 2^{k+1}\}.$$

The set $\mathcal{S}_k$ is encoded with a binary significance map $b_k[m]$:

$$b_k[m] = \begin{cases} 0 & \text{if } m \notin \mathcal{S}_k \\ 1 & \text{if } m \in \mathcal{S}_k \end{cases} . \tag{34}$$

An embedded algorithm quantizes $a[m]$ uniformly with a quantization step (bin size) $\Delta = 2^n$ that is progressively reduced. Let $m \in \mathcal{S}_k$ with $k \ge n$. The amplitude $|Q(a[m])|$ of the quantized number is represented in base 2 by a binary string with non-zero digits between the bit k and the bit n. The bit k is necessarily 1 because $2^k \le |Q(a[m])| < 2^{k+1}$. Hence, $k-n$ bits are sufficient to specify this amplitude, to which one bit is added for the sign.

The embedded coding is initiated with the largest quantization step to produce at least one non-zero quantized coefficient. To refine the quantization step from 2^{n+1} to 2^n, the algorithm records the significance map $b_n[m]$ and the sign of $a[m]$ for $m \in \mathcal{S}_n$. This can be done by directly recording the sign of significant coefficients with a variable incorporated into the significance map $b_n[m]$. Afterwards, the code stores the bit n of all amplitudes $|Q(a[m])|$ for $m \in \mathcal{S}_k$ with $k > n$. If necessary, the coding precision is improved by decreasing n and continuing the encoding. The different steps of the algorithm can be summarized as follows [12]:

1. Store the index n of the first non-empty set $\mathcal{S}_n$, where $n = \left\lfloor \sup_m \log_2 |a[m]| \right\rfloor$.
2. Store the significance map $b_n[m]$ and the sign of $a[m]$ for $m \in \mathcal{S}_n$.
3. Store the n^{th} bit of all coefficients $|a[m]| > 2^{n+1}$. These are coefficients that belong to some set $\mathcal{S}_k$ for $k > n$, whose coordinates were already stored. Their n^{th} bit is stored in the order in which their position was recorded in the previous passes.
4. Decrease n by 1 and go to step 2.

This algorithm may be stopped at any time in the loop, providing a code for any specified number of bits. The distortion rate is analyzed when the algorithm is stopped at step 4. All coefficients above $T = 2^n$ are uniformly quantized with a bin size $\Delta = 2^n$. The zero quantization bin $[-T, T]$ is therefore twice as big as the other quantization bins. This quantizer is the same as in the direct transform coding studied in Section 3.1, for a zero-bin ratio $\theta = \frac{T}{\Delta} = 1$. This value was shown to be nearly optimal for wavelet image coders. The total distortion $D = D_0 + D_1$ is therefore not modified by the embedding strategy.

Once the algorithm stops, we denote by M the number of significant coefficients above $T = 2^n$. The total number of bits of the embedded code is $R = R_0^e + R_1^e$, where R_0^e is the number of bits needed to encode all significance maps $b_k[m]$ for $k \geq n$, and R_1^e is the number of bits used to encode the amplitudes of the quantized significant coefficients $Q(a[m])$, knowing that $m \in \mathcal{S}_k$ for $k > n$.

To appreciate the efficiency of this embedding strategy, the bit budget $R_0^e + R_1^e$ is compared to the number of bits $R_0 + R_1$ used by the direct transform coder of Section 3.1. The value R_0 is the number of bits used to encode the overall significance map

$$b[m] = \begin{cases} 0 & \text{if } |a[m]| \leq T \\ 1 & \text{if } |a[m]| > T \end{cases} \tag{35}$$

and R_1 is the number of bits used to encode the quantized significant coefficients.

An embedded strategy encodes $Q(a[m])$ knowing that $m \in \mathcal{S}_k$ and hence that $2^k \leq |Q(a[m])| < 2^{k+1}$, whereas a direct transform coding knows only that $|Q(a[m])| > T = 2^n$. Thus fewer bits are needed for embedded codes: $R_1^e \leq R_1$. This improvement may be offset however by the supplement of bits needed to encode the significance maps $\{b_k[m]\}_{k>n}$ of the sets $\{\mathcal{S}_k\}_{k>n}$. A direct transform coder records a single significance map $b[m]$, which specifies $\cup_{k \geq n} \mathcal{S}_k$. It provides less information and is therefore encoded with fewer bits: $R_0^e \geq R_0$. An embedded coder brings an improvement over a direct transform coder if

$$R_0^e + R_1^e \leq R_0 + R_1.$$

This can happen if we have some prior information about the position of large decomposition coefficients $Q(a[m])$ versus smaller ones. It allows us to reduce the number of bits needed to encode the partial sorting of all coefficients provided by the significance maps $\{b_k[m]\}_{k>n}$. The use of such prior information produces an overhead of R_0^e relative to R_0 that is smaller than the gain of R_1^e relative to R_1. Let $x(z)$ be the sorted amplitudes of the coefficients $a[m]$. The following theorem computes R_1^e with the same hypotheses as Theorem 2, allowing us to compare it with R_1.

Theorem 4 *Suppose that $x(z)$ satisfies (19-22). Let $\gamma_M = \gamma(\frac{M}{N}) > \frac{1}{2}$. If $\frac{M}{N} \leq \epsilon$ and $M \geq \frac{1}{\epsilon}$ then*

$$r_1^e = \frac{R_1^e}{M} = 1 + \frac{1}{2^{\frac{1}{\gamma_M}} - 1} \Big[1 + O(\epsilon |\log_2 \epsilon|) \Big] \ , \tag{36}$$

and

$$D(\bar{R}) = (1 + K)\, D_0 \left(\frac{\bar{R}}{r_1^e + r_0^e} \right) , \tag{37}$$

where $r_0^e = \frac{R_0^e}{M}$ and $K = \frac{D_1}{D_0}$ is given by (24).

The proof of this theorem is in appendix A.3. In the following, we omit the corrective terms to simplify the notation. This theorem proves that $\frac{R_1^e}{M}$ is well approximated by

$$r_1^e = 1 + \frac{1}{2^{\frac{1}{\gamma_M}} - 1}. \tag{38}$$

Figure 12 verifies that the value of $\frac{R_1^e}{M} - 1$ calculated numerically with the embedded wavelet coding software of Said and Pearlman [12] is close to the estimate $r_1^e - 1$ calculated by computing γ_M from Figure 4(a). We subtract 1 bit because Said and Pearlman do not encode the sign bits with the amplitudes of the significant coefficients but store their values in the significance maps.

Let us compare $r_1^e = \frac{R_1^e}{M}$ calculated in (36) with the value $r_1 = \frac{R_1}{M}$ estimated in (25) for a direct transform coding, with $\theta = 1$. The bit budget of an embedded coding is smaller than that of a direct transform coding if $r_0^e + r_1^e \leq r_0 + r_1$ and hence

$$r_0^e - r_0 \leq r_1 - r_1^e = (1 + \gamma_M)\log_2 \mathrm{e} + \log_2 \gamma_M - \frac{1}{2^{\frac{1}{\gamma_M}} - 1}. \tag{39}$$

If $\gamma_M \approx 1$ then $r_1 - r_1^e \approx 1.9$. The inequality (39) is satisfied for embedded transform codings implemented in wavelet bases [13] and in a block cosine basis [17], by taking advantage of prior knowledge of the location of large versus small coefficients using zero-trees.

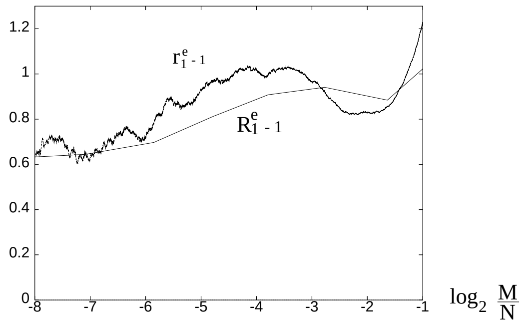


Figure 12: Comparison of $\frac{R_1^e}{M} - 1$ with the theoretical estimate $r_1^e - 1 = \frac{1}{2^{\frac{1}{\gamma_M}} - 1}$ for the embedded coding of Lena.

A wavelet coefficient $\langle f, \psi_{j,p,q}^k \rangle$ has a large amplitude where the signal has sharp transitions. If an image f is Lipschitz α in the neighborhood of (x_0, y_0), then for wavelets $\psi_{j,p,q}^k$ located in this neighborhood one can prove [11] that there exists $A \geq 0$ such that

$$|\langle f, \psi_{j,p,q}^k \rangle| \leq A\, 2^{j(\alpha+1)}.$$

The worst singularities are often discontinuities which means that $\alpha \geq 0$. In the neighborhood of singularities without oscillations, the amplitudes of wavelet coefficients thus decrease when the scale 2^j decreases. This property is not valid for oscillatory patterns. High frequency oscillations create coefficients at large scales 2^j that are typically smaller than those at the fine scale that matches the period of oscillation. Such oscillatory patterns are not often encountered in images although they do appear as thin lines in the Barbara image.

Wavelet zero-trees, introduced by Lewis and Knowles [9], take advantage of the decay of wavelet coefficients by relating these coefficients across scales with quad-trees. These zero-trees take advantage of a partial self-similarity of the image [2]. Shapiro [13] used this zero-tree structure to encode the embedded significance maps of wavelet coefficients. Numerical experiments were performed with Said and Pearlman's software [12], which improves Shapiro's zero-tree coder with a set partitioning technique. A cubic spline orthogonal wavelet basis was used. Figure 13 displays the value of $\frac{R}{M} = r_0^e + r_1^e$ as a function of $\log_2 \frac{M}{N}$ for the four test images. These curves have variations centered at 5.5. This graph should be compared with Figure 7 that shows $r_0 + r_1$ calculated with a direct wavelet image coder. An improvement of approximately 1 bit per significant coefficient is obtained.

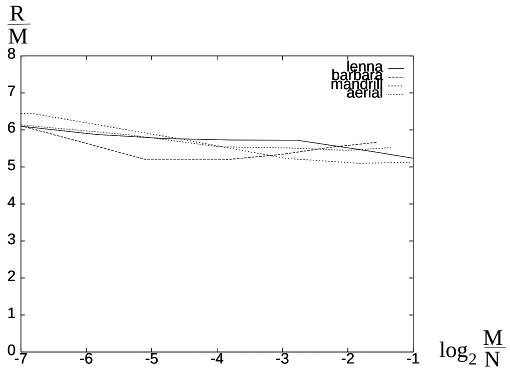


Figure 13: Variations of $\frac{R}{M}$ for an embedded wavelet coding of the test images.

The embedded distortion rate function is calculated with (37). Inserting $\gamma_M \approx 1$ in (24) yields $K \approx \frac{1}{12}$. Since $r_0^e + r_1^e \approx 5.5$, we get an approximate distortion rate formula

$$\hat{D}(\bar{R}) = (1 + \frac{1}{12})\, D_0 \left(\frac{\bar{R}}{5.5} \right) .$$

Figure 14 compares the $P_{SNR}(D)$ calculated numerically for the our test images and its theoretical approximation $P_{SNR}(\hat{D})$. Once more we verify that the distortion rate essentially depends only upon the approximation error function $D_0(z)$. The variations of the constants K and $r_0^e + r_1^e$ can be neglected. The embedding strategy reduces $r_0 + r_1$ to $r_0^e + r_1^e$, but the variations of the distortion rate still depends on the non-linear approximation error $D_0(z)$.

5 Conclusion

We have shown that at low bit rates, the distortion rate of transform coders can be computed by separating the coefficients quantized to zero from all others. The resulting distortion rate $D(\bar{R})$

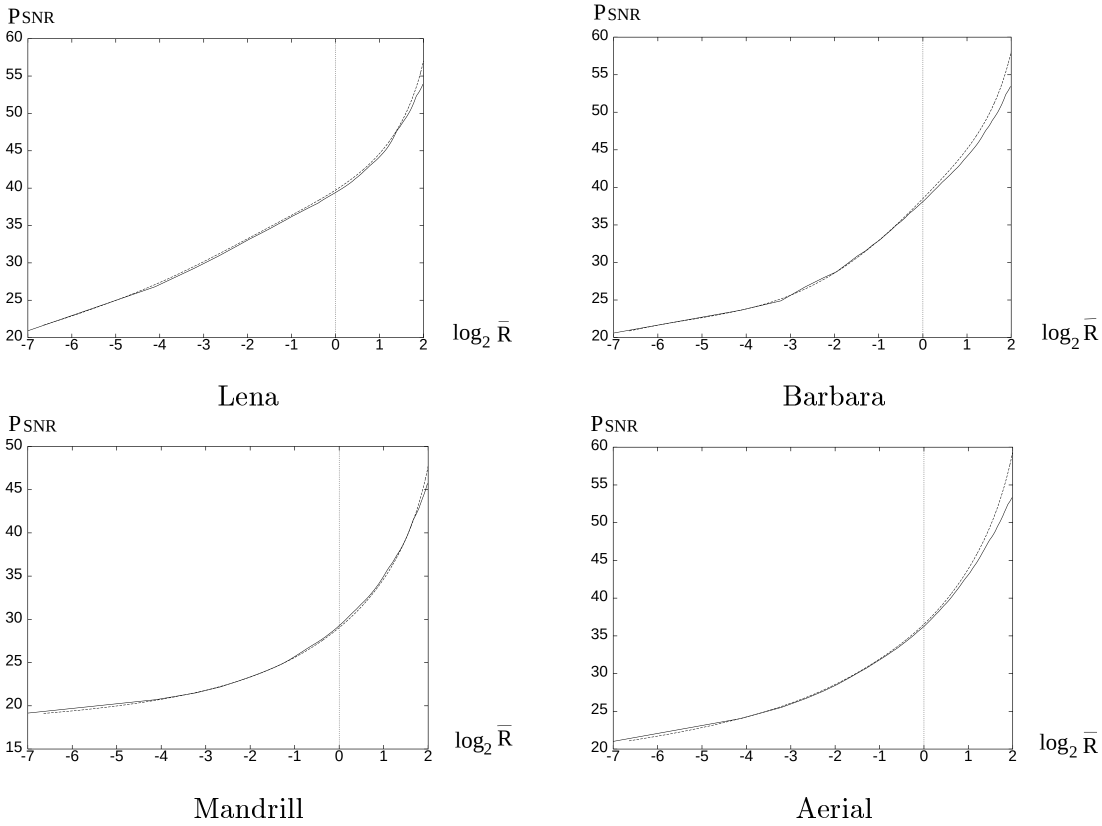


Figure 14: $P_{SNR}(D)$ (solid line) and $P_{SNR}(\hat{D})$ (dashed line) for an embedded wavelet coding, with $\hat{D} = (1 + \frac{1}{12})\, D_0(\frac{\bar{R}}{5.5})$.

depends crucially on the precision of non-linear image approximations with few non-zero basis coefficients. In wavelet and block cosine bases, we have demonstrated that if $\bar{R} \leq 1$ then $D(\bar{R})$ decays like $C\,\bar{R}^{1-2\gamma}$ where γ is an exponent of the order of 1, which varies slowly as a function of $\log_2 \bar{R}$. Embedded transform coders improve the constant C but not the exponent γ, which depends on the image.

A Proofs of Theorems

A.1 Proof of Theorem 2

The distortion rate formula (23) is derived by observing that $D = D_0 + D_1$ and $\bar{R} = \frac{R_0}{N} + \frac{R_1}{N}$. By definition $K = \frac{D_1}{D_0}$, $r_0 = \frac{R_0}{M}$ and $r_1 = \frac{R_1}{M}$. Inserting these variables in the equation $D = D_0 + D_1$ yields (23).

The main difficulty is in computing $K = \frac{D_1}{D_0}$ and $r_1 = \frac{R_1}{M}$. For this purpose, $x(z)$ is approximated by a function $x_M(z)$ that is tangential to $x(z)$ at $z = \frac{M}{N}$. We first estimate K and r_1 by replacing $x(z)$ with $x_M(z)$ and then evaluate the error introduced by this approximation. We define

$$x_M(z) = T\left(\frac{N\ z}{M}\right)^{-\gamma_M}.$$

Since there are exactly M coefficients above T, $x(\frac{M}{N}) = T = x_M(\frac{M}{N})$ and

$$\frac{d\log_2 x(\frac{M}{N})}{d\log_2 z} = \frac{d\log_2 x_M(\frac{M}{N})}{d\log_2 z} = -\gamma_M.$$

Both curves are thus tangential at $z = \frac{M}{N}$, and the concavity of $\log_2 x(z)$ guarantees that

$$\log_2 x(z) \leq \log_2 x_M(z).$$

Let us now prove that $\frac{D_1}{D_0}$ is given by (24). We compute

$$D_0 = \sum_{k=M}^{N-1} x^2(\frac{k}{N}) \approx N\int_{\frac{M}{N}}^{1} x^2(z)dz. \tag{40}$$

This value is approximated by

$$\begin{aligned}\hat{D}_0 &= N\int_{\frac{M}{N}}^{1} x_M^2(z)dz = T^2 N\int_{\frac{M}{N}}^{1}\left(\frac{Nz}{M}\right)^{-2\gamma_M} dz \\ &= \frac{T^2 M}{2\gamma_M - 1}\left(1 + \left(\frac{M}{N}\right)^{2\gamma_M - 1}\right).\end{aligned} \tag{41}$$

The concavity of $\log_2 x(z)$ implies that $x(z) \leq x_M(z)$ for $z \geq \frac{M}{N}$ and hence that $\hat{D}_0 - D_0 \geq 0$. To compute $\hat{D}_0 - D_0$ we write

$$x(z) = T\left(\frac{Nz}{M}\right)^{-\beta(z)}, \tag{42}$$

with

$$\beta(z) = \frac{\log_2 x(\frac{M}{N}) - \log_2 x(z)}{\log_2 z - \log_2 \frac{M}{N}}.$$

The estimation error is

$$\hat{D}_0 - D_0 = NT^2 \int_{\frac{M}{N}}^{1} \left(\frac{Nz}{M}\right)^{-2\gamma_M} \left[1 - \left(\frac{Nz}{M}\right)^{-2(\beta(z)-\gamma_M)}\right] dz.$$

We cut this integral in two parts

$$|\hat{D}_0 - D_0| \leq I_1 + I_2, \tag{43}$$

with

$$I_1 = NT^2 \int_{\frac{M}{N}}^{\frac{1}{2\epsilon}\frac{M}{N}} \left(\frac{Nz}{M}\right)^{-2\gamma_M} \left[1 - \left(\frac{Nz}{M}\right)^{-2(\beta(z)-\gamma_M)}\right] dz \tag{44}$$

and

$$I_2 = NT^2 \int_{\frac{1}{2\epsilon}\frac{M}{N}}^{+\infty} \left(\frac{Nz}{M}\right)^{-2\gamma_M} dz = \frac{MT^2(2\epsilon)^{2\gamma_M - 1}}{2\gamma_M - 1}. \tag{45}$$

To find an upper bound of I_1 we compute an upper bound of $\beta(z) - \gamma_M$. Clearly

$$\beta\left(\frac{M}{N}\right) = \gamma\left(\frac{M}{N}\right) = \gamma_M.$$

Since $\log_2 x(z)$ is concave, $|\gamma_M - \beta(z)| \leq |\gamma_M - \gamma(z)|$. The hypothesis (19) guarantees that

$$\forall z \in [0, 2^{-1}] \quad , \quad \left|\frac{d\gamma(z)}{d\log_2 z}\right| \leq \epsilon,$$

and hence

$$|\gamma_M - \gamma(z)| \ \leq \ \epsilon \left|\log_2 z - \log_2 \frac{M}{N}\right| \ , \tag{46}$$

which yields

$$\forall z \in [0, 2^{-1}] \quad , \quad |\gamma_M - \beta(z)| \ \leq \epsilon \left|\log_2 \frac{Nz}{M}\right|. \tag{47}$$

The concavity of $\log_2 x(z)$ also implies that $\beta(z) - \gamma_M \geq 0$ for $z \geq \frac{M}{N}$. Since $\frac{1}{2\epsilon}\frac{M}{N} \leq 2^{-1}$ it follows that

$$\begin{aligned}
I_1 \ &\leq \ NT^2 \int_{\frac{M}{N}}^{\frac{1}{2\epsilon}\frac{M}{N}} \left(\frac{Nz}{M}\right)^{-2\gamma_M} \left[1 - \left(\frac{Nz}{M}\right)^{-2\epsilon\log_2 \frac{Nz}{M}}\right] dz \\
&\leq \ NT^2 \int_{\frac{M}{N}}^{\frac{1}{2\epsilon}\frac{M}{N}} \left(\frac{Nz}{M}\right)^{-2\gamma_M} \left[1 - 2^{-2\epsilon\left(\log_2 \frac{Nz}{M}\right)^2}\right] dz \\
&\leq \ MT^2 \int_{1}^{\frac{1}{2\epsilon}} z^{-2\gamma_M} \left[1 - 2^{-2\epsilon(\log_2 z)^2}\right] dz.
\end{aligned}$$

One can then derive that

$$I_1 \leq \frac{M\,T^2}{2\gamma_M - 1}\, \mathcal{O}\left(\epsilon|\log_2 \epsilon|^2\right). \tag{48}$$

Inserting (41,45,48) in (43) proves (24).

Let us now prove that $\frac{R_1}{M}$ satisfies (25). To compute R_1, we must calculate the differential entropy

$$\mathcal{H}_d(X_T) = -\int_{-\infty}^{+\infty} p_T(x) \log_2 p_T(x) dx,$$

where $p_T(x) = \frac{N}{M}\ p(x)\ \mathbf{1}_{\{|x|>T\}}(x)$. Since $p(x) = p(-x)$

$$\mathcal{H}_d(X_T) = -2\,\frac{N}{M} \int_{T}^{+\infty} p(x) \log_2 p(x) dx. \tag{49}$$

This integral is calculated by relating it to $x(z)$. Since

$$z(x) = 1 - \int_{-x}^{x} p(u) du, \tag{50}$$

it follows that $z'(x) = -p(x) - p(-x) = -2p(x)$ and hence

$$x'(z) = \frac{1}{z'(x)} = \frac{-1}{2p(x)}. \tag{51}$$

Since $p(x)dx = -\frac{1}{2}dz$, the change of variable $z = z(x)$ in (49) yields

$$\mathcal{H}_d(X_T) = 2\frac{N}{M} \int_0^{\frac{M}{N}} \log_2 |2x'(z)|\ \frac{dz}{2},$$

and with a change of variable

$$\mathcal{H}_d(X_T) = 1 + \int_0^1 \log_2 \left(\frac{M}{N} \left| x' \left(\frac{M}{N} z \right) \right| \right) dz. \tag{52}$$

This integral is first estimated by replacing $x(z)$ by $x_M(z)$

$$\begin{aligned}
\hat{\mathcal{H}}_d &= 1 + \int_0^1 \log_2 \left(\frac{M}{N} \left| x'_M \left(\frac{M}{N}\, z \right) \right| \right)\ dz \\
&= 1 + \int_0^1 \log_2 (T \gamma_M z^{-\gamma_M - 1})\, dz \\
&= 1 + \log_2 T + \log_2 \gamma_M + (\gamma_M + 1) \log_e 2.
\end{aligned} \tag{53}$$

Let us compute the error

$$|\hat{\mathcal{H}}_d - \mathcal{H}_d(X_T)| \leq \left| \int_0^1 \log_2 \left(\frac{x'(\frac{M}{N} z)}{x'_M(\frac{M}{N} z)} \right)\ dz \right|.$$

Observe that

$$\gamma(z) = -\frac{d \log_2 x(z)}{d \log_2 z} = \frac{-z x'(z)}{x(z)}$$

and $\gamma_M = -\frac{r x'_M(z)}{x_M(z)}$ so

$$\begin{aligned}
|\hat{\mathcal{H}}_d - \mathcal{H}_d(X_T)| &\leq \int_0^1 \left| \log_2 \left(\frac{\gamma(\frac{M}{N} z)}{\gamma_M} \right) + \log_2 \left(\frac{x_M(\frac{M}{N} z)}{x(\frac{M}{N} z)} \right) \right|\ dz \\
&\leq \int_0^1 \left| \log_2 \left(\frac{\gamma(\frac{M}{N} z)}{\gamma_M} \right) + \left(\gamma_M - \beta(\frac{M}{N} z) \right)\ \log_2 z \right|\ dz.
\end{aligned} \tag{54}$$

Since $\inf_{z\in[0,1]} \gamma(z) = \mu > 0$ it follows that

$$\left|\log_2\left(\frac{\gamma(\frac{M}{N}z)}{\gamma_M}\right)\right| = \left|\log_2\left(1+\frac{\gamma_M-\gamma(\frac{M}{N}z)}{\gamma(\frac{M}{N}z)}\right)\right| \leq \frac{1}{\mu}\left|\gamma\left(\frac{M}{N}z\right)-\gamma_M\right|.$$

Since $\frac{M}{N}z \leq 2^{-1}$, we proved in (46,47) that

$$\left|\gamma\left(\frac{M}{N}z\right)-\gamma_M\right| \leq \epsilon\,|\log_2 z| \quad \text{and} \quad \left|\gamma_M-\beta\left(\frac{M}{N}z\right)\right| \leq \epsilon\,|\log_2 z|.$$

By inserting these inequalities in (54) one can verify that $|\hat{\mathcal{H}}_d - \mathcal{H}_d(X_T)| = \mathcal{O}(\epsilon)$. From (53) and $T=\theta\Delta$ we thus derive that

$$\frac{R_1}{M} = \mathcal{H}_d(X_T) - \log_2\Delta = \log_2\theta + 1 + (1+\gamma_M)\log_2 \mathrm{e} + \log_2\gamma_M + \mathcal{O}(\epsilon)\,,$$

which finishes the proof of (25).

Let us finally prove (26). We must calculate

$$\frac{d\log_2 D_0(\frac{M}{N})}{d\log_2 z} = \frac{dD_0(\frac{M}{N})}{dz}\,\frac{M}{ND_0(\frac{M}{N})}. \tag{55}$$

We derive from (40) that

$$\frac{dD_0(\frac{M}{N})}{dz} = -Nx^2(\frac{M}{N}) = -NT^2$$

(24) yields

$$\frac{1}{D_0} = \frac{2\gamma_M-1}{MT^2}\left[1+\mathcal{O}\left(\epsilon|\log_2\epsilon|^2+\epsilon^{2\gamma_M-1}\right)\right].$$

Inserting these last two equations in (55) gives (26).

A.2 Proof of Theorem 3

We proved in (24) that

$$D = D_1\left(1+\frac{12\,\theta^2}{2\gamma_M-1}\right)$$

Since $D_1 = \frac{M\Delta^2}{12} = \frac{MT^2}{12\theta^2}$ and $T = x(\frac{M}{N})$

$$\begin{aligned} D &= M\,T^2\left(\frac{1}{\theta^2}+\frac{12}{2\gamma_M-1}\right) \\ &= M\,x^2\left(\frac{M}{N}\right)\left(\frac{1}{\theta^2}+\frac{12}{2\gamma_M-1}\right). \end{aligned}$$

We decompose $\frac{R}{M} = \frac{R_0}{M} + \frac{R_1}{M}$ where $\frac{R_0}{M} = r_0$ is a constant and (25) shows that

$$\frac{R_1}{M} = r_1(M,\theta) = 1 + (1+\gamma_M)\log_2\mathrm{e} + \log_2\gamma_M + \log_2\theta + \mathcal{O}(\epsilon|\log_2\epsilon|)\,. \tag{56}$$

If we neglect the residual terms

$$\frac{R}{M} = 1 + r_0 + (1+\gamma_M)\log_2\mathrm{e} + \log_2\gamma_M + \log_2\theta.$$

The variable M depends on (R,θ) and we can thus write $\frac{R}{M} = \beta(R,\theta)$. Hence

$$D(\bar{R},\theta) = \frac{N\bar{R}}{\beta(R,\theta)}\; x^2\left(\frac{\bar{R}}{\beta(R,\theta)}\right)\;\left(\frac{1}{\theta^2} + \frac{12}{2\gamma_M - 1}\right). \tag{57}$$

To compute $\frac{\partial D(\bar{R},\theta)}{\partial\theta}$ we need to calculate $\frac{\partial\beta(R,\theta)}{\partial\theta}$. Observe that

$$\frac{\partial\beta(R,\theta)}{\partial\theta} = \frac{\partial\frac{R}{M}}{\partial\theta} = -\frac{R}{M^2}\frac{\partial M(R,\theta)}{\partial\theta}.$$

We compute $\frac{\partial M(R,\theta)}{\partial\theta}$ by taking the derivative with respect to θ at R fixed of the equality

$$R = M\Big[1 + r_0 + (1+\gamma_M)\log_2 \mathrm{e} + \log_2\gamma_M + \log_2\theta\Big].$$

We get

$$0 = \frac{\partial M(R,\theta)}{\partial\theta}\Big[\frac{R}{M} + M\log_2 \mathrm{e}\frac{d\gamma_M}{dM} + M\frac{1}{\gamma_M}\frac{d\gamma_M}{dM}\Big] + M\frac{\log_2 \mathrm{e}}{\theta}.$$

The slow variation condition (19) on the slope imposes that

$$\left|\frac{d\gamma_M}{d\log_2(\frac{M}{N})}\right| \leq \epsilon.$$

Inserting this in the previous equation proves that

$$\frac{\partial M(R,\theta)}{\partial\theta}\Big[\frac{R}{M} + \mathcal{O}(\epsilon)\Big] = -M\,\frac{\log_2 \mathrm{e}}{\theta}$$

Hence

$$\frac{\partial\beta(R,\theta)}{\partial\theta} = -\frac{R}{M^2}\,\frac{\partial M(R,\theta)}{\partial\theta} = \frac{\log_2 \mathrm{e}}{\theta}\Big[1 + \mathcal{O}(\epsilon)\Big] = \beta'. \tag{58}$$

Let us now compute a derivative from (57)

$$\frac{\partial\log_{\mathrm{e}} D(R,\theta)}{\partial\theta} = -\frac{\beta'}{\beta(R,\theta)} + 2\,\left(-\frac{R\beta'}{\beta^2(R,\theta)}\right)\frac{x'(\frac{M}{N})}{Nx(\frac{M}{N})} - \frac{\frac{2}{\theta^3}}{\frac{1}{\theta^2} + \frac{12}{2\gamma_M - 1}} = 0. \tag{59}$$

We know that

$$\frac{d\log_2 x(\frac{M}{N})}{d\log_2(\frac{M}{N})} = \frac{Mx'(\frac{M}{N})}{Nx(\frac{M}{N})} = -\gamma_M. \tag{60}$$

Inserting (58) and (60) with $\beta(R,\theta) = \frac{R}{M} = r_1 + r_0$ in (59) proves (30).

A.3 Proof of Theorem 4

Suppose that the algorithm stops at n. It performs a quantization with intervals of size $T = 2^n$. For each coefficient $2^k \leq |a[m]| < 2^{k+1}$, we saw that the number of bits required to specify the quantized value of $|a[m]|$ is

$$p = k - n = \left\lfloor \log_2\frac{|a[m]|}{T}\right\rfloor,$$

and we need one bit for the sign of $a[m]$. The coefficients above 2^n are the M coefficients whose amplitudes are given by $x(z)$. So

$$R_1^e = \sum_{m=1}^{M} \left\lfloor \log_2 \frac{x(\frac{m}{N})}{T} \right\rfloor + M = I_1 + I_2 + M \tag{61}$$

where I_1 corresponds to the sum for $m \geq \epsilon M \geq 1$

$$I_1 = \sum_{m=M\epsilon}^{M} \left\lfloor \log_2 \left(\frac{M}{m}\right)^{\beta(\frac{m}{N})} \right\rfloor,$$

and I_2 is calculated from $m < \epsilon M$

$$I_2 = \sum_{m=1}^{M\epsilon - 1} \left\lfloor \log_2 \left(\frac{M}{m}\right)^{\beta(\frac{m}{N})} \right\rfloor.$$

We proved in (47) that

$$\left|\gamma_M - \beta\left(\frac{m}{N}\right)\right| \leq \epsilon \, \left|\log_2\left(\frac{m}{M}\right)\right|. \tag{62}$$

As a consequence, for $m \in [\epsilon M, M]$

$$\gamma_M - \epsilon \, |\log_2 \epsilon| = \beta_0 \leq \beta\left(\frac{m}{N}\right) \leq \beta_1 = \gamma_M + \epsilon \, |\log_2 \epsilon|. \tag{63}$$

Let us define

$$I(\beta) = \sum_{m=M\epsilon}^{M} \left\lfloor \log_2 \left(\frac{M}{m}\right)^{\beta} \right\rfloor.$$

The inequalities (63) imply that

$$I(\beta_0) \leq I_1 \leq I(\beta_1). \tag{64}$$

Let us calculate $I(\beta)$ for any β. We decompose the sum in slices where

$$2^i \leq \left(\frac{M}{m}\right)^{\beta} < 2^{i+1},$$

for which the floor of the $\log_2$ is equal to i. It corresponds to

$$M 2^{\frac{-(i+1)}{\beta}} < m \leq M 2^{\frac{-i}{\beta}}.$$

Let $a = |\log_2 \epsilon|$. We obtain

$$I(\beta) = \sum_{i=0}^{a\beta} i \, M \left(2^{\frac{-i}{\beta}} - 2^{\frac{-(i+1)}{\beta}}\right) \tag{65}$$

so

$$\frac{I(\beta)}{M} = \left(1 - 2^{-\frac{1}{\beta}}\right) \sum_{i=0}^{a\beta} i \, 2^{-\frac{i}{\beta}}. \tag{66}$$

The reader can verify that

$$\sum_{i=0}^{+\infty} i 2^{-\frac{i}{\beta}} = \frac{2^{-\frac{1}{\beta}}}{\left(1 - 2^{-\frac{1}{\beta}}\right)^2}. \tag{67}$$

Approximating the finite sum (65) by this infinite sum yields

$$\frac{I(\beta)}{M} = \frac{1}{2^{\frac{1}{\beta}} - 1} + \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right). \tag{68}$$

The inequality (64) can thus be rewritten

$$\frac{1}{2^{\frac{1}{\beta_0}} - 1} - \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right) \leq \frac{I_1}{M} \leq \frac{1}{2^{\frac{1}{\beta_1}} - 1} + \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right).$$

Since $\beta_1 - \gamma_M = \gamma_M - \beta_0 = \epsilon \,|\log_2 \epsilon|$ it follows that

$$\left|\frac{I_1}{M} - \frac{1}{2^{\frac{1}{\gamma_M}} - 1}\right| \leq \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right). \tag{69}$$

To compute the discrete sum I_2 we use (62) which proves that

$$\beta\left(\frac{m}{N}\right) \leq \gamma_M + \epsilon \left|\log_2 \frac{m}{M}\right|.$$

Hence

$$\begin{aligned} I_2 &= \sum_{m=1}^{M\epsilon - 1} \left\lfloor \log_2 \left(\frac{M}{m}\right)^{\gamma_M + \epsilon |\log_2 \frac{m}{M}|} \right\rfloor \\ &\leq M \int_0^{\epsilon} (\gamma_M + \epsilon |\log_2 z|)\, |\log_2 z| \; dz = M\, \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right). \end{aligned}$$

We thus derive from (61) and (69) that

$$\left|\frac{R_1^e}{M} - \frac{1}{2^{\frac{1}{\gamma_M}} - 1} - 1\right| = \mathcal{O}\left(\epsilon \,|\log_2 \epsilon|\right).$$

References

[1] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard. On the importance of combining wavelet-based non-linear approximation with coding strategies, 1997. Preprint.

[2] G. M. Davis. A wavelet-based analysis of fractal image compression. *IEEE Trans. Image Proc.*, 1997.

[3] R. A. DeVore, B. Jawerth, and B. J. Lucier. Image compression through wavelet transform coding. *IEEE Trans. Info. Theory*, 38(2):719–746, March 1992.

[4] N. Farvardin and J. Modestino. Optimum quantizer performances for a class of non-gaussian memoryless sources. *IEEE Trans. Info. Theory*, 30(3):485–497, May 1984.

[5] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992. Chapter 2.

[6] H. Gish and J. N. Pierce. Asymptotically efficient quantizing. *IEEE Trans. Info. Theory*, 14:676–683, September 1968.

[7] N. J. Jayant and P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, Englewood-Cliffs, NJ, 1984.

[8] E. LePennec. Modélisation d'images par ondelettes, December 1997. Mémoire de DEA, CMAP, Ecole Polytechnique, Paris, France.

[9] A. S. Lewis and G. Knowles. Image compression using the 2-D wavelet transform. *IEEE Trans. Image Proc.*, 1(2):244–250, April 1992.

[10] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Patt. Recog. and Mach. Intell.*, 11(7):674–693, July 1989.

[11] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, New York, 1998.

[12] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circ. and Syst. for Video Tech.*, 6(3):243–250, June 1996.

[13] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Proc.*, 41(12):3445–3462, December 1993.

[14] G. Sullivan. Efficient scalar quantization of exponential and laplacian random variables. *IEEE Trans. Info. Theory*, 42(5):1365–1374, September 1996.

[15] G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, April 1991.

[16] I. Witten, R. Neal, and J. Cleary. Arithmetic coding for data compression. *Comm. of the ACM*, 30(6):519–540, 1987.

[17] Z. X. Xiong, O. Guleryuz, and M. T. Orchard. Embedded image coding based on DCT. In *VCIP in Europeen Image Proc. Conf.*, 1997.