HAL
open science

A Bit Allocation Method for Sparse Source Coding

Mounir Kaaniche, Aurélia Fraysse, Béatrice Pesquet-Popescu,
Jean-Christophe Pesquet

To cite this version:

Mounir Kaaniche, Aurélia Fraysse, Béatrice Pesquet-Popescu, Jean-Christophe Pesquet. A Bit Allocation Method for Sparse Source Coding. IEEE Transactions on Image Processing, 2014, 23 (1), pp.137-152. hal-00796891v4

HAL Id: hal-00796891
https://hal.science/hal-00796891v4

Submitted on 25 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Bit Allocation Method for Sparse Source Coding

Mounir Kaaniche, *Member IEEE*, Aurélia Fraysse, *Member IEEE*,
Béatrice Pesquet-Popescu, *Fellow IEEE* and Jean-Christophe Pesquet, *Fellow IEEE*

***Abstract*—In this paper, we develop an efficient bit allocation strategy for subband-based image coding systems. More specifically, our objective is to design a new optimization algorithm based on a rate-distortion optimality criterion. To this end, we consider the uniform scalar quantization of a class of mixed distributed sources following a Bernoulli-Generalized Gaussian distribution. This model appears to be particularly well-adapted for image data which have a sparse representation in a wavelet basis. In this paper, we propose new approximations of the entropy and the distortion functions by using piecewise affine and exponential forms, respectively. Thanks to these approximations, bit allocation is reformulated as a convex optimization problem. Solving the resulting problem allows us to derive the optimal quantization step for each subband. Experimental results show the benefits that can be drawn from the proposed bit allocation method in a typical transform-based coding application.**

***Index Terms*—Bit allocation, sparse sources, generalized Gaussian, lossy source coding, rate-distortion theory, piecewise approximation, convex optimization.**

I. Introduction

In image and video coding systems, it is desired to achieve the best possible image quality for a given bitrate or, conversely, to minimize the bitrate for a given image quality. To this respect, a great attention has been paid to the problem of bit allocation where a given amount of bits must be efficiently distributed among blocks of a DCT-coded image or among subbands of a wavelet-based coder [2], or among frames in a video sequence [3]. The general framework behind the bit allocation strategy is Rate-Distortion (R-D) theory which aims at minimizing the average distortion of the input signal subject to a constraint on the available global bitrate. Since both the rate and the distortion measures in a typical transform coding scheme are controlled by the choice of the quantizers, the major issue is to find the optimal quantization steps for the constrained minimization problem. It is thus necessary to study the rate and distortion functions of the source to be encoded. Two main classes of methods have been developed to deal with the bit allocation problem: numerical- and analytical-based approaches. Algorithms in the first category aim at empirically estimating the R-D curves and resort to some iterative techniques to find the optimal quantization parameters [3]. For instance, Lagrangian optimization techniques have been well investigated in the literature [4], [5], [6]. In these approaches, the constrained minimization problem is transformed into an unconstrained version by incorporating the constraint into the objective function. In [4], a bit allocation method for completely arbitrary input signals (or blocks) and discrete quantizer sets is considered in the case of independent coding contexts. An extension of this work to subband coding has been proposed in [2]. Another extension to a dependent coding environment has also been considered in [6]. More precisely, the authors describe the R-D Lagrangian cost function in the form of a trellis and use the Viterbi algorithm to find the optimal solution for coders exploiting temporal and spatial dependencies such as MPEG and pyramidal coders. In [7], the bit allocation problem is converted into the graph theoretic problem of finding the shortest path in a directed acyclic graph. Besides, it should be noticed that dynamic programming algorithms [3], [8] and descent algorithms [9], [10] have also been proposed to select the optimal quantization parameters.

It is important to note that these numerical methods may be computationally intensive since a large number of R-D operating points must be measured for each subband in order to obtain R-D curves which are both differentiable and convex [11]. In other words, the R-D data are first evaluated for all possible quantization settings. Then, the optimal solution is derived. For instance, the EBCOT algorithm of JPEG2000 relies on the explicit computation of several truncation points, for each code-block of size 32×32 or 64×64 of the subband coefficients, and a post-optimization by selecting the best R-D points for each Lagrangian parameter λ. Generally, a larger number of truncation points per code-block induces a higher complexity, but also better optimization results. In order to reduce the complexity, André *et al.* [12] have recently proposed to perform the computation of a few points (i.e. for some possible quantization settings) and interpolate them using spline approximations.

Thus, to further overcome the complexity of these numerical methods, alternative approaches which do not require the estimation of R-D curves have also been developed. These approaches provide closed-form expressions of the R-D function by assuming various input distributions and quantizer characteristics. For instance, the performance of optimum scalar quantizers subject to an entropy constraint was investigated through numerical methods [13], [14] for different source probability densities (e.g uniform, Gaussian, Laplacian, Generalized Gaussian) at low resolution (i.e. bitrate). In [15],

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

M. Kaaniche is with Université Paris 13, Laboratoire de Traitement et Transport de l'Information, Institut Galilée, 93430 Villetaneuse, France. E-mail: mounir.kaaniche@univ-paris13.fr.

A. Fraysse is with Université Paris-Sud, Supélec, Laboratoire des Signaux et Systèmes, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France. E-mail: fraysse@lss.supelec.fr .

B. Pesquet-Popescu is with Télécom ParisTech, Signal and Image Processing Department, 46 rue Barrault, 75014 Paris, France. E-mail: pesquet@telecom-paristech.fr.

J.-C. Pesquet is with Université Paris-Est, Laboratoire d'Informatique Gaspard Monge and CNRS UMR 8049, 77454 Marne-la-Vallée, France. E-mail: pesquet@univ-mlv.fr.

Part of this work has been presented in [1].

a parametric representation of the operational R-D function of a scalar quantizer is derived for a uniformly distributed source and a wide class of distortion measures. In [16], a distortion measure based on differential entropy has been introduced for image coding using uniform scalar quantization. In [17], an approach for designing entropy constrained scalar quantizers for exponential and Laplace distributions is proposed and comparisons are made with uniform quantizers. Recently in [18], the asymptotic behavior of a uniform quantizer is characterized at low resolution for a memoryless Gaussian source and a squared error distortion measure. Other studies have also considered the use of Laplace and Generalized Gaussian probability models in modern compression systems [19], [20], [21].

While using R-D models leads to a reduction of the complexity of the optimal quantization parameter selection for the different subbands in the context of still image compression, classical high rate R-D models [22] play also a crucial role in the selection of the modes, for example in the context of H.264/AVC video coding standard [23]. The slope of the R-D curve is directly related to the quantization parameter of the macroblock, and shown to be equal to the Lagrangian parameter for the mode selection. Furthermore, in order to estimate the motion vector for each macroblock, an R-D optimization is performed for the different partition modes, and the partition leading to the smallest R-D cost is selected. Due to complexity reasons, the mode and partition optimization is performed recursively from macroblock to macroblock, thus leading to a (possibly) not optimal global solution. Moreover, in a multi-layer coding environment, an R-D model in the ρ-domain has been developed [24] based on similar high rate approximations.

The main contribution of this paper is to design an efficient bit allocation algorithm in a subband coding context (typically, for wavelet-based coders) by adopting an analytical approach. More precisely, we will consider the uniform scalar quantization of the different subband coefficients resulting from a multiresolution analysis. Moreover, in order to provide a general framework, we develop the theoretical part by considering a Bernoulli-Generalized Gaussian (BGG) model which was found to be well-suited for modelling sparse wavelet coefficients [25], [26]. It must be emphasized that the developed theoretical results remain also valid for standard images which are often modeled by using a Generalized Gaussian (GG) distribution [27], [28]. After extending recent approximation formulas for the entropy and the distortion of uniformly quantized BGG sources [29], we propose a piecewise affine (resp. piecewise exponential) form of the entropy (resp. distortion) which allows us to get fine low rate and high rate approximations of these functions. Thanks to the proposed approximations, we are able to reformulate the bit allocation problem by making use of convex analysis tools. Following this approach, we derive explicit expressions of the optimal quantization parameters of the subbands. Finally, simulations are performed on natural images which can be simply modeled by a GG distribution, and then on images with flat regions (like the cartoon ones) where a BGG model appears to be more appropriate than the former one.

The remainder of this paper is organized as follows: in Section II, we define the probabilistic model for the considered subband coefficients as well as the quantizer characteristics. We introduce the resulting entropy and distortion functions. In Section III, we provide new piecewise convex approximations of the entropy and the distortion. In Section IV, we reformulate the bit allocation problem as a set of convex optimization problems, for which we derive the optimal solutions. Finally, an application of the proposed method to transform-based image coding is illustrated in Section V and some conclusions are drawn in Section VI.

II. Entropy and distortion of a uniformly quantized BGG source

A. Source and quantization models

First, we consider the problem of coding an input signal by performing a wavelet (or frame-based) decomposition. Let us assume that the source to be quantized is composed of J subbands having n_j coefficients ($j \in \{1, \dots, J\}$) so that $n = \sum_{j=1}^{J} n_j$ is the total number of coefficients. Since the empirical distribution of the detail coefficients shows a very high number of small amplitude (close to zero) values, an appropriate way for characterizing their sparsity in the j-th subband consists of modelling these coefficients with a BGG distribution whose probability density function f_j is defined by:

$$\forall \xi \in \mathbb{R}, \qquad f_j(\xi) = (1-\epsilon_j)\delta(\xi) + \epsilon_j \tilde{f}_j(\xi) \tag{1}$$

where $\epsilon_j \in [0,1]$ is a mixture parameter, δ denotes the Dirac distribution (i.e. point mass at 0) and $\tilde{f}_j$ is the probability density function for a GG distribution with shape parameter $\beta_j \in]0,2]$ and scale factor $\omega_j \in]0,+\infty[$:

$$\forall \xi \in \mathbb{R}, \qquad \tilde{f}_j(\xi) = \frac{\beta_j \omega_j^{1/\beta_j}}{2\Gamma(1/\beta_j)} e^{-\omega_j |\xi|^{\beta_j}} \tag{2}$$

where Γ is the gamma function. Recall that the differential entropy of such a GG variable is given [30] as:

$$h_{\beta_j}(\omega_j) = -\int_{-\infty}^{\infty} \tilde{f}_j(\xi) \log_2 \tilde{f}_j(\xi)\, d\xi = \log_2\left(\frac{2\Gamma(1/\beta_j)}{\beta_j \omega_j^{1/\beta_j}}\right) + \frac{1}{\beta_j}.$$

Each coefficient $X_{j,s}$ with $s \in \{1, \dots, n_j\}$ in subband $j \in \{1, \dots, J\}$ is quantized before being entropy coded. For this purpose, we assume that, for each subband j, a scalar uniform quantizer with a quantization step q_j and having a deadzone of size $(2\tau_j - 1)q_j$ where $\tau_j > 1/2$ is used [31]. Note that $\tau_j = 1$ corresponds to a deadzone of size q_j. Thus, for every $s \in \{1, \dots, n_j\}$, the output $\overline{X}_{j,s}$ of the quantizer is given by:

$$\overline{X}_{j,s} = r_0 = 0, \quad \text{if } |X_{j,s}| < \left(\tau_j - \tfrac{1}{2}\right) q_j, \quad \text{where } \tau_j > 1/2$$

and, for all $i \in \mathbb{Z}$, $\overline{X}_{j,s} = r_{i,j}$,

$$(\text{if } (\tau_j + i - \tfrac{3}{2})q_j \leq X_{j,s} < (\tau_j + i - \tfrac{1}{2})q_j \text{ and } i \geq 1)$$
$$\text{or } (\text{if } (-\tau_j + i + \tfrac{1}{2})q_j < X_{j,s} \leq (-\tau_j + i + \tfrac{3}{2})q_j \text{ and } i \leq -1),$$

where the reconstruction levels are given by

$$\forall i \geq 1, \qquad r_{i,j} = -r_{-i,j} = (\tau_j + i - 1 + \zeta_j)q_j \tag{3}$$

and $\zeta_j \in [-1/2, 1/2]$ is an "offset" parameter indicating the shift of the reconstruction level with respect to the center of the quantization interval. Note that we will not consider any saturation effect. The most commonly used quantization rule corresponds to the case when $\zeta_j = 0$ (i.e. mid-point reconstruction). For example, this rule constitutes the basic ingredient of many encoding strategies (such as EBCOT [32]) which have been developed in wavelet-based image compression techniques.

Since the objective of the paper is to focus on the bit allocation problem for the quantized coefficients, it is now necessary to study their rate and distortion functions.

B. Entropy and distortion measures

As frequently done in the development of R-D algorithms, we approximate the bitrate of a memoryless source by the zero-order entropy of the quantized coefficients [22], [33]. Thus, by assuming that the random variable $X_{j,s}$ with $j \in \{1, \ldots, J\}$ and $s \in \{1, \ldots, n_j\}$ is distributed according to (1), the entropy of the associated quantized variable $\overline{X}_{j,s}$ with $j \in \{1, \ldots, J\}$ and $s \in \{1, \ldots, n_j\}$ is given by:

$$H_{f_j}(q_j, \epsilon_j) = -\sum_{i=-\infty}^{\infty} \mathsf{p}_{i,j} \log_2 \mathsf{p}_{i,j} \tag{4}$$

where, for every $i \in \mathbb{Z}$, $\mathsf{p}_{i,j} = \mathsf{P}(\overline{X}_{j,s} = r_{i,j})$ represents the probability of occurrence of the $r_{i,j}$ reconstruction level.

We also propose to express the distortion function by using the p_j-th order moment of the quantization error:

$$e_j(q_j, \epsilon_j) = \mathsf{E}[|X_{j,s} - \overline{X}_{j,s}|^{p_j}] \tag{5}$$

where $p_j \geq 1$ is a real exponent. In particular, $p_j = 2$ corresponds to the mean square error criterion whereas $p_j = 1$ corresponds to the mean absolute one. Taking real values of the exponent which depend on j provides flexibility in the choice of the distortion measure.

It is important to note here that close approximations of the entropy and asymptotic expressions of the distortion of a quantized BGG random variable are provided in [29] for both low and high bitrates. However, these approximations have been derived in the case of log-concave distributions (more precisely when $1 \leq \beta_j \leq 2$) and for a quantizer with a deadzone of size q_j (i.e. $\tau_j = 1$). It is worth pointing that in practice, typical values of β_j can be smaller than 1 and the size of the deadzone can be parameterized to have a different value for each subband (as in JPEG2000 Part 2, while in Part 1, a typical deadzone of size $2q_j$ is used) [34]. Therefore, the main approximation results given in [29] need to be extended in this paper by incorporating a nontrivial deadzone in the quantizer and also considering the case when $\beta_j < 1$.

Once the entropy and distortion functions have been defined, the bit allocation problem can be formulated. In our case, this problem consists of finding the quantization steps for each subband or, equivalently, the vector $\mathbf{q} = (q_1, q_2, \ldots, q_J) \in [0, +\infty[^J$ minimizing the average distortion

$$D(\mathbf{q}) = \sum_{j=1}^{J} \rho_j e_j(q_j, \epsilon_j) \tag{6}$$

where $\forall j \in \{1, \ldots, J\}, \quad \rho_j \in]0, +\infty[$ and $\sum_{j=1}^{J} \rho_j = 1,$

subject to the constraint that the total bitrate is equal to or smaller than a target bitrate $R_{\max}$:

$$H(\mathbf{q}) = \sum_{j=1}^{J} \frac{n_j}{n} H_{f_j}(q_j, \epsilon_j) \leq R_{\max}. \tag{7}$$

Note that, for orthonormal representations, when for every $j \in \{1, \ldots, J\}$, $p_j = 2$ and $\rho_j = n_j/n$, $D(\mathbf{q})$ is also equal to the distortion in the spatial domain. For other scenarios (biorthogonal representations or redundant frames), a good approximation of the distortion in the spatial domain can be obtained in a number of cases by appropriate choices of the constants $(\rho_j)_{1 \leq j \leq J}$ [35]. The degrees of freedom in the choices of the constants $(p_j)_{1 \leq j \leq J}$ and $(\rho_j)_{1 \leq j \leq J}$ can also be exploited in order to define perceptual criteria [36] better fitting the Human Visual System (HVS) characteristics.

III. Approximations of the entropy and of the distortion

The objective of this section is to develop accurate approximations of the entropy and the distortion for a general BGG source model. These approximations will allow us to reformulate the bit allocation problem in a more tractable form.

A. Piecewise affine approximation of the entropy

Let Q_a with $a \in \mathbb{R}_+^*$ be the normalized incomplete Gamma function [37], defined as

$$\forall \xi \in \mathbb{R}, \qquad Q_a(\xi) = \frac{1}{\Gamma(a)} \int_0^{\xi} \theta^{a-1} e^{-\theta} d\theta. \tag{8}$$

A close approximation of the entropy of a quantized BGG source can be obtained as follows:

Proposition 1: For a quantized BGG random variable distributed according to (1)*, the entropy $H_{f_j}(q_j, \epsilon_j)$ can be approximated by*

$$\widehat{H}_{f_j}(q_j, \epsilon_j) = \Phi(\mathsf{p}_{0,j}, \epsilon_j) + \epsilon_j \widehat{H}_{\widetilde{f}_j}(q_j) \tag{9}$$

with $$\Phi(\mathsf{p}_{0,j}, \epsilon_j) = -\big(1 - \epsilon_j(1 - \mathsf{p}_{0,j})\big) \log_2 \big(1 - \epsilon_j(1 - \mathsf{p}_{0,j})\big) - \epsilon_j(1 - \mathsf{p}_{0,j}) \log_2 \epsilon_j + \epsilon_j \mathsf{p}_{0,j} \log_2 \mathsf{p}_{0,j},$$

and $$\begin{aligned} \widehat{H}_{\widetilde{f}_j}(q_j) = & -\mathsf{p}_{0,j} \log_2 \mathsf{p}_{0,j} - 2\mathsf{p}_{1,j} \log_2 \mathsf{p}_{1,j} \\ & + \big(h_{\beta_j}(\omega_j) - \log_2 q_j\big)\Big(1 - Q_{1/\beta_j}\Big(\omega_j \big(\tau_j + \frac{1}{2}\big)^{\beta_j} q_j^{\beta_j}\Big)\Big) \\ & + \frac{\omega_j^{1/\beta_j}(\tau_j + \frac{1}{2}) q_j}{\Gamma(1/\beta_j)} e^{-\omega_j (\tau_j + \frac{1}{2})^{\beta_j} q_j^{\beta_j}}. \end{aligned} \tag{10}$$

The error incurred in this approximation is such that

$$0 \leq \widehat{H}_{f_j}(q_j, \epsilon_j) - H_{f_j}(q_j, \epsilon_j) \leq 2\epsilon_j q_j C(\beta_j, \tau_j) \tilde{f}_j\big((\tau_j + \frac{1}{2}) q_j\big),$$

$$\textit{with} \quad C(\beta_j, \tau_j) = \begin{cases} \left(\frac{2\tau_j + 1}{2\tau_j - 1}\right)^{1-\beta_j} & \textit{if} \quad \beta_j < 1 \\ \left(\frac{2\tau_j + 2}{2\tau_j + 1}\right)^{\beta_j - 1} & \textit{if} \quad \beta_j \in [1, 2]. \end{cases} \tag{11}$$

Proof: See Appendix A.

It is worth pointing out that such an approximation formula may be useful in practice in the sense that it allows us to efficiently compute the entropy for any given set of quantization steps.

Generally, analytical-based R-D algorithms use the standard Bennett formula to obtain a close approximation of the entropy [29], [22]. This high-resolution approximation formula, which is also valid when a quantizer with a deadzone is used and $\beta_j \in (0,2]$, allows us to express the entropy of the j-th subband as an affine function of $l_j = \log_2(q_j)$:

$$H_{f_j}(q_j, \epsilon_j) = H_{\epsilon_j} + \epsilon_j(h_{\beta_j}(\omega_j) - l_j) + o(l_j 2^{l_j}) \qquad (12)$$

where $H_{\epsilon_j} = -\epsilon_j \log_2 \epsilon_j - (1-\epsilon_j)\log_2(1-\epsilon_j)$ is the entropy of a Bernoulli random variable with parameters $(1-\epsilon_j, \epsilon_j)$.

However, the approximation formula (10) is not tractable for optimization purposes, whereas (12) is only valid at high resolution (i.e. when q_j is small). In order to develop a bit allocation strategy well-adapted for both high and low resolutions, we propose to define a piecewise convex approximation of the entropy function by considering a more flexible function of $\mathbf{l} = (l_1, l_2, \ldots, l_J)$, given by $\sum_{j=1}^{J} \frac{n_j}{n} g_j(l_j)$, where g_j have the following piecewise affine form:

$$\forall j \in \{1,\ldots,J\}, \;\; g_j(l_j) = a_j^k l_j + c_j^k \;\; \text{if } l_j^{(h,k-1)} \leq l_j \leq l_j^{(h,k)} \qquad (13)$$

with $k \in \{1, 2, \ldots, m^{(h)}\}$ and $m^{(h)}$ is a given parameter corresponding to the considered number of intervals (i.e. the number of segments chosen to approximate the entropy). For every $j \in \{1,\ldots,J\}$, the parameters $(a_j^k)_{1\leq k\leq m^{(h)}}$ are nonpositive reals, and the parameters $(c_j^k)_{1\leq k\leq m^{(h)}}$ are real numbers. Note that the superscript h has been used to distinguish between the intervals used for the approximation of the entropy and those later used for the approximation of the distortion.

B. Practical implementation of the entropy approximation

In practice, we set $l_j^{(h,0)} = -\infty$ and we choose the other points $\left(l_j^{(h,k)}\right)_{1\leq k\leq m^{(h)}}$ in such a way that the resulting piecewise affine function constitutes a good approximation of the entropy H_{f_j} of the source. More precisely, for the first interval, the high resolution approximation (12) can be employed, leading to

$$\forall\, j \in \{1,\ldots,J\}, \qquad a_j^1 = -\epsilon_j \quad \text{and} \quad c_j^1 = H_{\epsilon_j} + \epsilon_j h_{\beta_j}(\omega_j)$$

By considering an arbitrary point $\tilde{l}_j^{(h,1)}$, we derive (a_j^2, c_j^2) such that g_j on the second interval is tangent to the graph of the entropy function $\widehat{H}_{f_j}$ at $\tilde{l}_j^{(h,1)}$. The computation of the tangent thus relies on the approximation of the entropy derived in Proposition 1. The upper bound of the first interval $l_j^{(h,1)}$, which corresponds also to the lower bound of the second interval, is then fixed to the abscissis of the intersection of the lines obtained on the first and second intervals. As explained below, the point $\tilde{l}_j^{(h,1)}$ is defined so as to get a small approximation error between the entropy $\widehat{H}_{f_j}$ and the function g_j on the first interval $[l_j^{(h,0)}, l_j^{(h,1)}]$. In this way, we obtain an approximation of the entropy function by using two intervals. By repeating the process, we compute the remaining values $(\tilde{l}_j^{(h,k)})_{2\leq k\leq m^{(h)}-1}$ which allow us to deduce the constants $(a_j^k)_{3\leq k\leq m^{(h)}}$, $(c_j^k)_{3\leq k\leq m^{(h)}}$ and $(l_j^{(h,k)})_{2\leq k\leq m^{(h)}-1}$. Since the entropy must be a nonnegative function, the last interval bound $l_j^{(h,m^{(h)})}$ is found such that $a_j^{m^{(h)}} l_j^{(h,m^{(h)})} + c_j^{m^{(h)}} = 0$. This entails:

$$\forall j \in \{1,\ldots,J\}, \quad g_j(l_j) = 0 \qquad \text{if } l_j \geq l_j^{(h,m^{(h)})}. \qquad (14)$$

It is worth pointing out that, in practice, the points $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$ (where the tangent to the graph of the entropy is computed) are determined only once (in an off-line computation step) by considering the possible values for $\beta_j \in]0,2]$ and $\epsilon_j \in [0,1]$. Indeed, we have observed from our tests that the points $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$ have similar values for the wavelet coefficients whose β_j values belong to a specific range $[\beta_j^{\min}, \beta_j^{\max}]$. Based on this observation, we have constructed look-up tables, for the selection of $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$, depending on intervals for the (ϵ_j, β_j) pair values. In other words, for a given interval for ϵ_j values, for $\beta_j \in]1.75, 2]$, we determine the optimal values of $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$. Then, for $\beta_j \in]1.5, 1.75]$, we determine the optimal values of $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$. After that, we proceed similarly for $\beta_j \in]1.3, 1.5]$, and so on.

Therefore, it is enough to estimate the distribution parameters per subband of the input image, and then to use the off-line computed values of $(\tilde{l}_j^{(h,k)})_{1\leq k\leq m^{(h)}-1}$ to deduce all the constants used in the piecewise convex approximation (13). From this viewpoint, our approach can be applied in real applications to any input image.

Fig. 1 illustrates the approximations of the entropy using two intervals ($m^{(h)} = 2$) and four intervals ($m^{(h)} = 4$). As expected, increasing the number of intervals leads to a better approximation of the entropy.

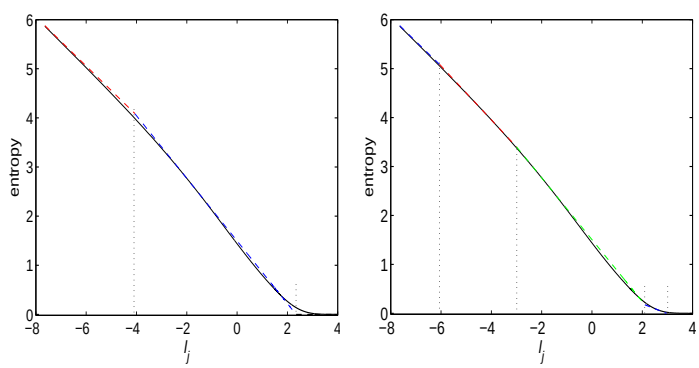


Fig. 1. Approximations g_j (in dashed line) of the entropy H_{f_j} (in solid line) of a uniformly quantized BGG source versus l_j: $m^{(h)} = 2$ (left side), $m^{(h)} = 4$ (right side). The parameters of the BGG source are $\epsilon_j = 0.5$, $\beta_j = 1.2$ and $\omega_j = 1$.

C. Piecewise exponential approximation of the distortion

On the other hand, a good approximation of the distortion of a quantized BGG source can be obtained as follows:

Proposition 2: For a quantized BGG random variable distributed according to (1)*, the distortion* $e_j(q_j, \epsilon_j)$ *can be*

approximated by

$$\begin{aligned}
&\widehat{e}_j(q_j,\epsilon_j)\\
&= 2\epsilon_j\Big(\frac{\omega_j^{-p_j/\beta_j}\Gamma((p_j+1)/\beta_j)}{2\Gamma(1/\beta_j)}Q_{(p_j+1)/\beta_j}\big(\omega_j(\tau_j-\frac{1}{2})^{\beta_j}q_j^{\beta_j}\big)\\
&+\int_{(\tau_j-\frac{1}{2})q_j}^{(\tau_j+\frac{1}{2})q_j}|\xi-r_{1,j}|^{p_j}\widetilde{f}_j(\xi)d\xi\\
&+\frac{\nu_j q_j^{p_j}}{2(p_j+1)}\big(1-Q_{1/\beta_j}\big(\omega_j(\tau_j+\frac{1}{2})^{\beta_j}q_j^{\beta_j}\big)\big)\Big) \qquad (15)
\end{aligned}$$

where the approximation error is such that

$$|e_j(q_j,\epsilon_j)-\widehat{e}_j(q_j,\epsilon_j)|\le 2\epsilon_j\frac{\nu_j q_j^{p_j+1}}{p_j+1}\widetilde{f}_j\big((\tau_j+\frac{1}{2})q_j\big). \quad (16)$$

Proof: See Appendix B.

Some comments can be made about this result:

- When $q_j\to 0$, the classical high resolution approximation is recovered:
$$e_j(q_j,\epsilon_j)=\epsilon_j\frac{\nu_j}{p_j+1}q_j^{p_j}(1+O(q_j)) \qquad (17)$$
where $\nu_j=(\frac{1}{2}+\zeta_j)^{p_j+1}+(\frac{1}{2}-\zeta_j)^{p_j+1}$.
- When $p_j=2$ (or more generally when p_j is an even integer), the integral in (15) can be easily expressed by using incomplete Gamma functions.

Similarly to the approximation of the entropy, Proposition 2 will be useful to compute both fast and accurate approximations of the distortion, but the derived expressions remain too intricate for developing efficient bit allocation algorithms.

We thus propose to use a rougher approximation of the distortion. More specifically, we propose to express the global distortion as a function of $\mathbf{l}=(l_1,l_2,\ldots,l_J)=\big(\log_2(q_1),\ldots,\log_2(q_J)\big)$ under the form $\sum_{j=1}^J\rho_j d_j(l_j)$, where d_j has the following piecewise exponential form:

$$\forall\, j\in\{1,\ldots,J\},$$
$$d_j(l_j)=\begin{cases}\epsilon_j(\alpha_j^k 2^{l_j\gamma_j^k}+\delta_j^k) & \text{if } l_j^{(d,k-1)}\le l_j<l_j^{(d,k)}\\ \epsilon_j\omega_j^{-p_j/\beta}\dfrac{\Gamma\big((p_j+1)/\beta_j\big)}{\Gamma(1/\beta_j)} & \text{if } l_j\ge l_j^{(d,m^{(d)})}\end{cases} \qquad (18)$$

where $k\in\{1,2,\ldots,m^{(d)}\}$ and $m^{(d)}$ is a given integer corresponding to the number of intervals used in our approximation. For every $j\in\{1,\ldots,J\}$, the parameters $(\alpha_j^k)_{1\le k\le m^{(d)}}$ and $(\gamma_j^k)_{1\le k\le m^{(d)}}$ are nonnegative reals, and the parameters $(\delta_j^k)_{1\le k\le m^{(d)}}$ are real numbers. While setting $\gamma_j^1=p_j$ and $\gamma_j^k=1$ for every $k\in\{2,\ldots,m^{(d)}\}$, and similarly to the selection procedure used with the approximation of the entropy (explained in Section III-B), the constants $(\alpha_j^k)_{1\le k\le m^{(d)}}$, $(\delta_j^k)_{1\le k\le m^{(d)}}$ and $(l_j^{(d,k)})_{0\le k\le m^{(d)}}$ (with $l_j^{(d,0)}=-\infty$) are determined in such a way that $d_j(l_j)$ constitutes a good approximation of $\widehat{e}_j(2^{l_j},\epsilon_j)$. In particular, by taking $\alpha_j^1=\frac{\nu_j}{p_j+1}$ and $\delta_j^1=0$, we obtain the high bitrate approximation of the distortion (see (17)) on the first interval $[l_j^{(d,0)},l_j^{(d,1)}]$.

Fig. 2 shows the approximations of the distortion for 2 and 4 intervals. It can be observed that setting $m^{(d)}$ to 2 results in a less precise approximation of the distortion e_j than $m^{(d)}=4$, especially at low bitrate. It can also be noticed from Figs. 1 and 2 that the chosen approximation interval bounds $l_j^{(h,1)}$, $l_j^{(h,2)}$ and $l_j^{(h,3)}$ for the entropy differ from those $l_j^{(d,1)}$, $l_j^{(d,2)}$ and $l_j^{(d,3)}$ for the distortion. This illustrates the fact that the selection steps for $\left(l_j^{(h,k)}\right)_{1\le k\le m^{(h)}}$ and $\left(l_j^{(d,k)}\right)_{1\le k\le m^{(d)}}$ should be performed independently in order to obtain good approximations of both the entropy and the distortion functions.

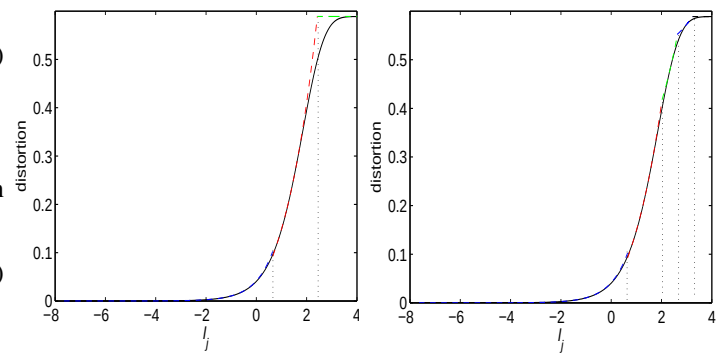


Fig. 2. Approximations d_j (in dashed line) of the distortion e_j (in solid line) of a uniformly quantized BGG source versus l_j: $m^{(d)}=2$ (left side), $m^{(d)}=4$ (right side). The parameters of the BGG source are $\epsilon_j=0.5$, $\beta_j=1.2$ and $\omega_j=1$.

IV. Proposed bit allocation method

In this part, we show how the approximations of the entropy and distortion functions proposed in the previous section allow us to solve the bit allocation problem in an efficient manner.

A. Optimization problem

Using the approximations g_j (resp. d_j) of the entropy in (13) (resp. of the distortion in (18)), the bit allocation problem defined at the end of Section II, can be recast as follows:

Problem 1: Find $\tilde{\mathbf{l}}$ minimizing the distortion function

$$\forall\, \mathbf{l}=(l_1,\ldots,l_J)\in\mathbb{R}^J,\qquad D(\mathbf{l})=\sum_{j=1}^J\rho_j d_j(l_j)$$

over the set C defined as

$$C:=\{\mathbf{l}=(l_1,\ldots,l_J)\in\mathbb{R}^J\mid\sum_{j=1}^J\frac{n_j}{n}g_j(l_j)\le R_{\max}\}. \quad (19)$$

A major difficulty for solving this problem stems from the fact that the functions g_j and d_j are non-differentiable and nonconvex. To define the different domains where the optimization is performed, we shall jointly sort the coefficients $\left(l_j^{(h,k)}\right)_{1\le k\le m^{(h)}}$ and $\left(l_j^{(d,k)}\right)_{1\le k\le m^{(d)}}$ in ascending order for each $j\in\{1,\ldots,J\}$. The resulting sorted coefficients will be denoted by $(l_j^1,\ldots,l_j^m)$ such that $l_j^1\le l_j^2\le\ldots\le l_j^m$ where $m\le m^{(h)}+m^{(d)}$. From the definition of the total bitrate constraint, a necessary condition for $\mathbf{l}$ to belong to C is

$$\forall\, j\in\{1,\ldots,J\},\quad a_j^1 l_j+c_j^1\le n n_j^{-1}R_{\max}. \qquad (20)$$

This means that, for every $j \in \{1,\ldots,J\}$, we can set the lower bound l_j^0 of the search interval to

$$l_j^0 = \min\Big(\frac{n n_j^{-1} R_{\max} - c_j^1}{a_j^1}, l_j^1\Big). \tag{21}$$

Moreover, since $g_j(l_j) = 0$ for every $l_j \geq l_j^m$, and d_j is an increasing function of l_j, it is clear that the optimal value of l_j will be lower than or equal to l_j^m. As a result, the problem is equivalent to minimize the distortion over the domain $[l_1^0, l_1^m] \times \ldots \times [l_J^0, l_J^m]$. In order to overcome the problem of the non-differentiability of the functions g_j or d_j at points $\big(l_j^k\big)_{1\leq k<m}$, we propose to subdivide the previous domain into boxes of the form $[l_1^{b_1}, l_1^{b_1+1}] \times \ldots \times [l_J^{b_J}, l_J^{b_J+1}]$ where $\mathbf{b} = (b_1,\ldots,b_J) \in \{0,\ldots,m-1\}^J$. On each box, the entropy and distortion are convex functions. Therefore, this subdivision technique leads to m^J subdomains where a convex optimization problem must be solved.

B. Solution of the bit allocation problem

In the following, we provide a closed form expression of the optimal quantization parameters. Suppose that $\mathbf{P_b} = [l_1^{b_1}, l_1^{b_1+1}] \times \ldots \times [l_J^{b_J}, l_J^{b_J+1}]$ corresponds to a given subdomain and let us denote by $(\mathcal{P}_\mathbf{b})$ the convex minimization problem on this subdomain. For concision purposes, let us introduce the following notation, for every $j \in \{1,\ldots,J\}$,

$$N_j = -\frac{n_j\, a_j^{b_j}}{\gamma_j^{b_j}}, \quad \kappa_j = \frac{n}{N_j}\rho_j \epsilon_j \alpha_j^{b_j} \ln 2, \tag{22}$$

$$\underline{\lambda}_j = \kappa_j 2^{\gamma_j^{b_j} l_j^{b_j}}, \quad \overline{\lambda}_j = \kappa_j 2^{\gamma_j^{b_j} l_j^{b_j+1}}. \tag{23}$$

The solution to the Problem $(\mathcal{P}_\mathbf{b})$ is given below.

Proposition 3:

(i) *If* $\sum_{j=1}^J \frac{n_j}{n} g_j(l_j^{b_{j+1}}) > R_{\max}$, *then there is no solution.*

(ii) *If* $\sum_{j=1}^J \frac{n_j}{n} g_j(l_j^{b_j}) \leq R_{\max}$, *then the solution is* $\widetilde{\mathbf{l}} = (l_1^{b_1},\ldots,l_J^{b_J})$.

(iii) *Otherwise, the solution is the vector* $\widetilde{\mathbf{l}}_\mathbf{b}$ *defined by*

$$\forall j \in \{1,\ldots,J\},\ \widetilde{l}_{j,\mathbf{b}} = \begin{cases} l_j^{b_j} & \text{if } j \in \mathbb{I} \\ \frac{1}{\gamma_j^{b_j}} \log_2\left(\frac{\widetilde{\lambda}}{\kappa_j}\right) & \text{if } j \in \mathbb{J} \\ l_j^{b_j+1} & \text{if } j \in \mathbb{K} \end{cases} \tag{24}$$

where

$$\widetilde{\lambda}^{N_\mathbb{J}} = \frac{2^{(\sum_{j=1}^J n_j c_j^{b_j} - n R_{\max})}}{2^{(\sum_{j\in\mathbb{I}} N_j \gamma_j^{b_j} l_j^{b_j} + \sum_{j\in\mathbb{K}} N_j \gamma_j^{b_j} l_j^{b_j+1})}} \prod_{j\in\mathbb{J}} \kappa_j^{N_j} \tag{25}$$

$$N_\mathbb{J} = \sum_{j\in\mathbb{J}} N_j \tag{26}$$

$$\mathbb{I} = \{j \in \{1,\ldots,J\} \mid \Phi'(\underline{\lambda}_j) \leq 0\}, \tag{27}$$

$$\mathbb{K} = \{j \in \{1,\ldots,J\} \mid \Phi'(\overline{\lambda}_j) > 0\} \tag{28}$$

$$\mathbb{J} = \{1,\ldots,J\} \setminus (\mathbb{I} \cup \mathbb{K}) \tag{29}$$

$$\forall\ \lambda \in \mathbb{R}_+,$$

$$\Phi(\lambda) = \lambda\Big(\sum_{j=1}^J \frac{n_j}{n} c_j^{b_j} - R_{\max}\Big) - \sum_{j=1}^J \varphi_j(\lambda) \tag{30}$$

with $\forall\ j \in \{1,\ldots,J\}$,

$$\varphi_j(\lambda) = \begin{cases} \frac{N_j}{n}(\gamma_j^{b_j} l_j^{b_j} \lambda - \frac{\underline{\lambda}_j}{\ln 2}) - \rho_j \epsilon_j \delta_j^{b_j} & \text{if } \lambda \leq \underline{\lambda}_j \\ \frac{N_j \lambda}{n \ln 2}(\ln(\frac{\lambda}{\kappa_j}) - 1) - \rho_j \epsilon_j \delta_j^{b_j} & \text{if } \underline{\lambda}_j < \lambda < \overline{\lambda}_j \\ \frac{N_j}{n}(\gamma_j^{b_j} l_j^{b_j+1} \lambda - \frac{\overline{\lambda}_j}{\ln 2}) - \rho_j \epsilon_j \delta_j^{b_j} & \text{if } \lambda \geq \overline{\lambda}_j. \end{cases}$$

Proof: See Appendix C.

The above expressions of the quantization parameters, obtained for each subdomain, allow us to determine a finite set of candidate distortion values. Once this has been performed, the subdomain leading to the global minimum distortion value is selected and its resulting quantization steps correspond to the optimal ones. It is worth pointing out that the computation of the quantization parameters as well as their corresponding distortion can be carried out for the subdomains independently of each other. Furthermore, it can be noticed that the maximum number m^J of these evaluations can be reduced by checking Conditions (i) and (ii) in Proposition 3.

V. Experimental results

In this part, we study the performance of the proposed bit allocation method in the context of transform-based coding applications. We employ the 9/7 biorthogonal wavelet transform, selected in the lossy coding mode of part I of the JPEG2000 compression standard. The decomposition is carried out over three resolution levels (i.e. $J = 10$) since the test image size (512×512) is not very large. Note also that the weights $(\rho_j)_{1\leq j\leq J}$ for the different wavelet subbands are computed by using the procedure presented in [38]. Our experiments have been performed for various standard test images[1] with different characteristics as shown in Fig. 3.

We should note that the wavelet transform is kept the same in all the experiments, and we are mainly interested in computing the optimal quantization parameters of the resulting wavelet subbands. As mentioned before, the first step of our method consists of modelling the resulting wavelet coefficients. For this purpose, we consider the two following models: the GG one and the more general BGG one.

A. GG-based model

In this case, the parameters β_j and ω_j for each subband are estimated directly from the corresponding coefficients, by using the maximum likelihood technique. Afterwards, we compute their corresponding entropy and distortion approximations and deduce their optimal quantization steps using Proposition 3. Figs. 4(a), 4(b) and 4(c) show the influence of the choice of the parameters $m^{(h)}$ and $m^{(d)}$ used for approximating the entropy and distortion functions. The plotted curve using the 'circle' symbols corresponds to the quadratic distortion (i.e. $p_j = 2$) resulting from an uniform scalar quantization of the GG model. The rate-distortion curve plotted using the 'star' symbol is obtained by performing a similar quantization of the wavelet coefficients of the image with the derived optimal quantizers. More precisely, we consider the cases $m^{(h)} = m^{(d)} = 2$, $m^{(h)} = m^{(d)} = 3$ and $m^{(h)} = m^{(d)} = 4$.

[1] http://sipi.usc.edu/database/
and http://homepages.cae.wisc.edu/~ece533/images/

It can be noticed that the difference between the plots corresponding to the theoretical GG source model and the image wavelet coefficients is reduced when the number of segments increases. In addition, one can observe from Fig. 4(d) that the image rate-distortion curves behave similarly when 3 or 4 approximating intervals are used. Based on this observation (which was confirmed by tests performed on other images), it can be concluded that there is no need to increase the number of segments, and therefore, it is sufficient in practice to use 3 or 4 intervals to approximate the entropy and distortion functions. Finally, we propose to compare the proposed bit allocation method with state-of-the-art methods based on Lagrangian optimization techniques [4]. More precisely, we consider the improved version of these methods, proposed recently in [12], where a spline interpolation method for rate-distortion curves is introduced. Fig. 5 shows the variations of the PSNR curves versus the entropy for different images. It can be observed that our method outperforms the state-of-the-art method by 0.2-1.2 dB. While the deadzone parameter τ_j is set to 1 in Fig. 5, Figs. 6(a) and 6(b) illustrate the performance of our method when the size of the deadzone is increased ($\tau_j = 2$). Thus, it can be noticed that the proposed method achieves a significant improvement compared with the state-of-the-art method.

B. BGG-based model

Although the GG model is well adapted to a large class of natural images, we have observed that this model is not the best suited for the class of images with *flat regions* separated by smooth contours. Examples of such images include cartoon ones and depth maps. To confirm this, we illustrate in Figs. 7(a) and 7(b) the histogram of the diagonal detail wavelet subband of the "cartoon" image at the first resolution level as well as the distribution used for modelling its coefficients. To find the best model, we propose to use a statistical goodness-of-fit test such as the Kolmogorov-Smirnov (KS) test which is based on the comparison of the cumulative distribution functions (cdf) [39]. Figs. 7(c) and 7(d) display these functions for both models with their resulting KS measure (the smallest measure value corresponds to the best fit). Hence, it can be noticed that the cdf associated with the BGG model is very close to the cdf associated with the subband wavelet coefficients. This illustrates the fact that the BGG model is more appropriate than the GG one for modelling very *sparse* representations. Based on this model, we have also employed the proposed bit allocation method for this class of images. Compared with the improved version of the Lagrangian based optimization technique [12], Figs. 8(a) and 8(b) show that the proposed method achieves an improvement of about 0.3-1 dB. In Fig. 8, the deadzone parameter τ_j is set to 1. Figs. 9(a) and 9(b) illustrate the performance of our method when the size of the deadzone is equal to $3q_j$ ($\tau_j = 2$). It can be concluded that the proposed method outperforms the state-of-the-art method in all these experiments. In addition, Fig. 10 shows the performance in terms of rate-distortion for the tested "cartoon" image when GG and BGG models are adopted. The obtained results confirm our previous observation concerning the interest of the statistical BGG model with respect to the GG one for very sparse representations.

Finally, in order to measure the relative gain of the proposed method, we used the Bjontegaard metric [40]. The results are illustrated in Table I for low and high bitrates corresponding respectively to the four bitrate points $\{0.1, 0.2, 0.3, 0.4\}$ and $\{0.7, 0.8, 0.9, 1\}$ bpp. Table I gives the gain of our method compared with the improved version of the Lagrangian based optimization technique [12]. Note that a bitrate saving with respect to the reference method corresponds to negative values. It can be observed that the proposed approach outperforms in average the state-of-the-art method by about -13 % and 0.7 dB in terms of entropy (bitrate saving) and PSNR. As it can be seen in Table I, the coding gain at low bitrate depends on the image, and it is close to that obtained at high bitrate for some images like "straw", "castle" and "cartoon". However, for other images, it can also be observed that the coding gains are higher at high bitrates than at low bitrates. This may be due to the fact that the resulting approximation error has more impact at low than at high bitrate.

For completeness sake, we have implemented our rate allocation scheme with two arithmetic entropy coders: one without context modeling (denoted by AC-1) and another one exploiting the previously encoded symbol as context (denoted by AC-2). The results, compared with JPEG2000 coder, are reported in Table II. We remark that our approach, followed by an arithmetic coder, can achieve better performance than JPEG2000 on some images ("straw" and "marseille"), while JPEG2000 does better on other images ("lena", "goldhill" and "einst"). The main reason for this is the high number of contexts used by JPEG2000 for the arithmetic coder, and this at each bitplane, not only on the final quantized coefficients. Moreover, due to the three passes per bitplane in the arithmetic coding, JPEG2000 performs actually a quantization with a fractional step size, which can be different in each code-block of size 32×32 or 64×64, while the current implementation of our method provides only one quantization step per subband.

Concerning the complexity of the proposed method, we first recall that the main steps behind our approach consist of computing the R-D curves for each subband and solving the convex optimization problem. Thus, thanks to Propositions 1 and 2, the convex approximations of the entropy and distortion functions are extremely fast. However, the solution of the convex optimization problem depends on the used number of intervals (since the maximum number of subdomains is equal to m^J as mentioned at the end of Section IV). For instance, for an image of size 512×512, and using a non-optimized Matlab implementation with an Intel Core 2 (2.93 GHz) computer, our optimization approach requires an execution time of about 2 and 30 seconds when 2 and 3 intervals are respectively employed, while the Lagrangian R-D optimization approach [12] takes about 15 seconds. When four or more intervals are used, the execution time becomes more important. It is worth pointing out that the computation of the quantization parameters as well as

their resulting distortions can be carried out in a *parallel* way for all the possible subdomains. Thus, much more efficient implementations on multicore architectures should significantly reduce these execution times.
All these results, obtained with different images, confirm the effectiveness of the considered probabilistic models and of the proposed bit allocation method.

VI. Conclusion

In this work, we have proposed to reformulate the bit allocation problem as a set of convex programming problems which can be dealt with in parallel. For this purpose, we have first proposed new piecewise convex approximations of the entropy and the distortion functions. Then, we have derived explicit expressions of the optimal quantization parameters which are valid in a given subdomain. This study has been carried out by considering two probabilistic models: the well-known GG model and its more general BGG form, which is particularly well-adapted for very *sparse* sources. Finally, we have illustrated through experimental results the benefits which can be drawn from the application of the proposed technique in transform-based coding application. The proposed approach can be further investigated in compression applications using other transforms, and the fine high rate and low rate approximations for the R-D curves could also be useful for the mode selection in video applications.
It is worth pointing out that the proposed scheme has been developed in the context of uniform scalar quantization which is retained in many embedded coders such as SPIHT, EBCOT and Tarp filter [41]. Furthermore, our approach can be used with different sizes of deadzone which is an interesting feature in JPEG2000 Part 2 where the deadzone can be parameterized to have a different width per subband. In our future work, we plan to extend this new promising scheme to various embedded coding systems. In addition, it should be noticed that our method can also be useful in the context of stereo image compression where the rate allocation process between the two views is not always well addressed in the literature.

Appendix A
Approximation of the entropy

We recall that the entropy of a quantized BGG random variable distributed according to (1) is given by [29]:

$$H_{f_j}(q_j,\epsilon_j) = \Phi(\mathsf{p}_{0,j},\epsilon_j) + \epsilon_j H_{\widetilde{f}_j}(q_j) \tag{31}$$

$$\text{where} \quad H_{\widetilde{f}_j}(q_j) = -\mathsf{p}_{0,j}\log_2 \mathsf{p}_{0,j} - 2\sum_{i=1}^{\infty} \mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} \tag{32}$$

is the entropy of a quantized GG random variable with probability density function $\widetilde{f}_j$. The probability of the zero level is

$$\mathsf{p}_{0,j} = 2\int_0^{q_j(\tau_j-\frac{1}{2})} \widetilde{f}_j(\xi)d\xi = Q_{1/\beta_j}\Big(\omega_j(\tau_j-\frac{1}{2})^{\beta_j} q_j^{\beta_j}\Big) \tag{33}$$

and the probability $\mathsf{p}_{i,j}$ of the $r_{i,j}$ reconstruction level, $i \geq 1$, is

$$\int_{(\tau_j+i-\frac{3}{2})q_j}^{(\tau_j+i-\frac{1}{2})q_j} \widetilde{f}_j(\xi)d\xi = \frac{1}{2}\Big(Q_{1/\beta_j}\Big(\omega_j\big((\tau_j+i-\frac{1}{2})q_j\big)^{\beta_j}\Big) - Q_{1/\beta_j}\Big(\omega_j\big((\tau_j+i-\frac{3}{2})q_j\big)^{\beta_j}\Big)\Big).$$

In the following, in order to prove the desired result, it is sufficient to show that the following approximation formula of the discrete entropy of a quantized GG random variable holds:

$$H_{\widetilde{f}_j}(q_j) = \widehat{H}_{\widetilde{f}_j}(q_j) + \Delta \tag{34}$$

$$\text{where} \quad 0 \leq \Delta \leq 2q_jC(\beta_j,\tau_j)\tilde{f}_j\big((\tau_j+\frac{1}{2})q_j\big). \tag{35}$$

Note that the case $\beta_j \in [1,2]$ was addressed in [29] for a quantizer with a deadzone of size q_j (i.e. $\tau_j = 1$). Let us now proceed to the general case.
Since $\widetilde{f}_j$ is a decreasing function on $\mathbb{R}_+$, we have

$$\forall\, i > 0, \quad q_j\widetilde{f}_j\big((\tau_j+i-1/2)q_j\big) \leq \mathsf{p}_{i,j} \leq q_j\widetilde{f}_j\big((\tau_j+i-3/2)q_j\big)$$

By noticing that

$$\begin{aligned} &-\mathsf{p}_{i,j}\log_2\mathsf{p}_{i,j} + \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\log_2\widetilde{f}_j(\xi)\,d\xi \\ &= \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)(\log_2\widetilde{f}_j(\xi) - \log_2\mathsf{p}_{i,j})\,d\xi \end{aligned} \tag{36}$$

we get the inequality:

$$\begin{aligned} &-\mathsf{p}_{i,j}\log_2\mathsf{p}_{i,j} + \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\log_2\widetilde{f}_j(\xi)\,d\xi \\ &\leq \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\Big(\log_2\widetilde{f}_j(\xi) - \log_2\widetilde{f}_j\big((\tau_j+i-1/2)q_j\big) \\ &\qquad - \log_2 q_j\Big)\,d\xi. \end{aligned} \tag{37}$$

On the other hand, from the positivity of the Kullback-Leibler divergence [42],

$$\int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \frac{\widetilde{f}_j(\xi)}{\mathsf{p}_{i,j}}\log_2\Big(\frac{\widetilde{f}_j(\xi)/\mathsf{p}_{i,j}}{1/q_j}\Big)\,d\xi \geq 0 \tag{38}$$

After developing (38) and using (37), we obtain for all $i \geq 1$

$$\begin{aligned} 0 \leq &-\mathsf{p}_{i,j}\log_2\mathsf{p}_{i,j} + \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\log_2\widetilde{f}_j(\xi)\,d\xi \\ &+ \log_2 q_j \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\,d\xi \\ &\leq \Big(\log_2\widetilde{f}_j((\tau_j+i-3/2)q_j) - \log_2\widetilde{f}_j\big((\tau_j+i-1/2)q_j\big)\Big) \\ &\times \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\,d\xi \\ &= \omega_j q_j^{\beta_j}\big((\tau_j+i-1/2)^{\beta_j} - (\tau_j+i-3/2)^{\beta_j}\big) \\ &\times \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\,d\xi. \end{aligned} \tag{39}$$

Now, two cases shall be considered:
• If $\beta_j < 1$, then, for every $i \geq 1$,

$$(\tau_j + i - 1/2)^{\beta_j} - (\tau_j + i - 3/2)^{\beta_j} \leq \beta_j(\tau_j + i - 3/2)^{\beta_j - 1},$$

where the upper bound follows from the fact that $\xi \mapsto \xi^{\beta_j}$ is a concave function when $\beta_j < 1$. In this case, we have

$$\begin{aligned}&\omega_j q_j^{\beta_j}\big((\tau_j + i - 1/2)^{\beta_j} - (\tau_j + i - 3/2)^{\beta_j}\big)\\ &\times \int_{(\tau_j+i-3/2)q_j}^{((\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\, d\xi\\ &\leq \beta_j\omega_j q_j\Big(q_j(\tau_j + i - \frac{1}{2}) - q_j\Big)^{\beta_j - 1} \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\, d\xi,\end{aligned}$$

$$\begin{aligned}\text{and } 0 \leq &-\mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} + \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi)\, d\xi\\ &+ \log_2 q_j \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\, d\xi\\ &\leq \beta_j\omega_j q_j \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} (\xi - q_j)^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi.\end{aligned}$$

It can be deduced that

$$\begin{aligned}0 \leq -\sum_{i=2}^{+\infty}\mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} &+ \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi) d\xi\\ &+ \log_2 q_j \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi) d\xi \leq I_1 \quad (40)\end{aligned}$$

$$\text{where} \qquad I_1 = \beta_j\omega_j q_j \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} (\xi - q_j)^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi.$$

Since $\xi \geq (\tau_j + 1/2)q_j \Leftrightarrow \xi - q_j \geq (2\tau_j - 1)\xi/(2\tau_j + 1)$, it can be concluded that

$$\begin{aligned}I_1 &\leq \beta_j\omega_j q_j\Big(\frac{2\tau_j - 1}{2\tau_j + 1}\Big)^{\beta_j-1} \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \xi^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi\\ &= \frac{\beta_j\omega_j^{1/\beta_j} q_j}{2\Gamma(1/\beta_j)}\Big(\frac{2\tau_j - 1}{2\tau_j + 1}\Big)^{\beta_j-1} e^{-\omega_j(\tau_j+\frac{1}{2})^{\beta_j} q_j^{\beta_j}}. \quad (41)\end{aligned}$$

• If $\beta_j \in [1, 2]$ then, for every $i \geq 1$,

$$\begin{aligned}&(\tau_j + i - 1/2)^{\beta_j} - (\tau_j + i - 3/2)^{\beta_j} = (\tau_j + i - 1)^{\beta_j}\\ &\times \Big((1 + \frac{1}{2(\tau_j + i - 1)})^{\beta_j} - (1 - \frac{1}{2(\tau_j + i - 1)})^{\beta_j}\Big)\\ &\leq \beta_j(\tau_j + i - 1)^{\beta_j-1}. \quad (42)\end{aligned}$$

Consequently,

$$\begin{aligned}&\omega_j q_j^{\beta_j}\big((\tau_j + i - 1/2)^{\beta_j} - (\tau_j + i - 3/2)^{\beta_j}\big) \int_{(\tau_j+i-\frac{3}{2})q_j}^{(\tau_j+i-\frac{1}{2})q_j} \widetilde{f}_j(\xi)\, d\xi\\ &\leq \beta_j\omega_j q_j\left(q_j(\tau_j + i - \frac{3}{2}) + \frac{q_j}{2}\right)^{\beta_j-1} \int_{(\tau_j+i-\frac{3}{2})q_j}^{(\tau_j+i-\frac{1}{2})q_j} \widetilde{f}_j(\xi)\, d\xi\end{aligned}$$

$$\begin{aligned}\text{and } 0 \leq &-\mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} + \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi)\, d\xi\\ &+ \log_2 q_j \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} \widetilde{f}_j(\xi)\, d\xi\\ &\leq \beta_j\omega_j q_j \int_{(\tau_j+i-3/2)q_j}^{(\tau_j+i-1/2)q_j} (\xi + \frac{q_j}{2})^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi.\end{aligned}$$

Thus,

$$\begin{aligned}0 \leq -\sum_{i=2}^{+\infty}\mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} &+ \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi) d\xi\\ &+ \log_2 q_j \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi) d\xi \leq I_2 \quad (43)\end{aligned}$$

$$\text{where} \qquad I_2 = \beta_j\omega_j q_j \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} (\xi + \frac{q_j}{2})^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi.$$

Since $\xi \geq (\tau_j + 1/2)q_j \Leftrightarrow \xi + \frac{q_j}{2} \leq (2\tau_j + 2)\xi/(2\tau_j + 1)$, it can be concluded that

$$\begin{aligned}I_2 &\leq \beta_j\omega_j q_j\Big(\frac{2\tau_j + 2}{2\tau_j + 1}\Big)^{\beta_j-1} \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \xi^{\beta_j-1}\widetilde{f}_j(\xi)\, d\xi\\ &= \frac{\beta_j\omega_j^{1/\beta_j} q_j}{2\Gamma(1/\beta_j)}\Big(\frac{2\tau_j + 2}{2\tau_j + 1}\Big)^{\beta_j-1} e^{-\omega_j(\tau_j+\frac{1}{2})^{\beta_j} q_j^{\beta_j}}. \quad (44)\end{aligned}$$

By combining (40) and (41) (resp. (43) and (44)) when $\beta_j < 1$ (resp. $\beta_j \in [1, 2]$), we get the following result:

$$\begin{aligned}0 \leq -\sum_{i=2}^{+\infty}\mathsf{p}_{i,j}\log_2 \mathsf{p}_{i,j} &+ \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi) d\xi\\ &+ \log_2 q_j \int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi) d\xi\\ &\leq \frac{\beta_j\omega_j^{1/\beta_j} q_j}{2\Gamma(1/\beta_j)} C(\beta_j, \tau_j) e^{-\omega_j(\tau_j+\frac{1}{2})^{\beta_j} q_j^{\beta_j}} \quad (45)\end{aligned}$$

where $C(\beta_j, \tau_j)$ is given by (11). Furthermore, it can be checked [29] that we have:

$$2\int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi) d\xi = 1 - Q_{1/\beta_j}\Big(\omega_j\big(\tau_j + \frac{1}{2}\big)^{\beta_j} q_j^{\beta_j}\Big) \quad (46)$$

$$\begin{aligned}\text{and} \quad &2\int_{(\tau_j+\frac{1}{2})q_j}^{+\infty} \widetilde{f}_j(\xi)\log_2 \widetilde{f}_j(\xi) d\xi\\ &= -h_{\beta_j}(\omega_j)\Big(1 - Q_{1/\beta_j}\Big(\omega_j\big(\tau_j + \frac{1}{2}\big)^{\beta_j} q_j^{\beta_j}\Big)\Big)\\ &\quad - \frac{\omega_j^{1/\beta_j}(\tau_j + \frac{1}{2})q_j}{\Gamma(1/\beta_j)}\, e^{-\omega_j(\tau_j+\frac{1}{2})^{\beta_j} q_j^{\beta_j}}. \quad (47)\end{aligned}$$

Therefore, the approximation formula of the entropy of the quantized GG random variable, given by (34)-(35), follows from (32), (45)-(47). Finally, the approximation formula for the discrete entropy of the quantized BGG random variable can be easily deduced from (31).

Concerning the high bitrate approximation of the entropy, it can be firstly noticed that $\Delta = O(q_j)$. We further know [43, p.891] that for all $a > 0$,

$$Q_a(\xi) = O(\xi^a), \qquad \text{as} \quad \xi \to 0. \quad (48)$$

Therefore, when $q_j \to 0$, we have

$$H_{\widetilde{f}_j}(q_j) = h_{\beta_j}(\omega_j) - \log_2 q_j + O(q_j). \tag{49}$$

Moreover, according to (33) and (48), we get

$$\begin{aligned}\Phi(\mathsf{p}_{0,j}, \epsilon_j) &= -\epsilon_j \log_2 \epsilon_j - (1-\epsilon_j)\log_2(1-\epsilon_j) \\ &\quad + \epsilon_j \mathbf{1}_{(0,1)}(\epsilon_j) \frac{\beta_j \omega_j^{1/\beta_j} q_j}{2\Gamma(1/\beta_j)} \log_2(\omega_j^{1/\beta_j} q_j) + O(q_j) \\ &= H_{\epsilon_j} + O(q_j \log_2 q_j) \end{aligned} \tag{50}$$

where $\mathbf{1}_{(0,1)}$ is the characteristic function of the interval $(0,1)$. Consequently, a high resolution approximation of the entropy of a quantized BGG random variable is given by (12).

Appendix B
Approximation of the distortion

If $X_{j,s}$ is distributed according to (1), the distortion evaluated through the p_j-th order moment of the quantization error is given by [29]:

$$\begin{aligned} e_j(q_j, \epsilon_j) = 2\epsilon_j \Big(&\int_0^{(\tau_j - \frac{1}{2})q_j} \xi^{p_j} \widetilde{f}_j(\xi) d\xi \\ &+ \sum_{i=1}^{+\infty} \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} \widetilde{f}_j(\xi) d\xi \Big). \end{aligned}$$

By noticing that

$$\begin{aligned} \int_0^{(\tau_j - \frac{1}{2})q_j} \xi^{p_j} \widetilde{f}_j(\xi) d\xi &= \frac{\omega_j^{-p_j/\beta_j} \Gamma((p_j+1)/\beta_j)}{2\Gamma(1/\beta_j)} \\ &\quad \times Q_{(p_j+1)/\beta_j}\big(\omega_j(\tau_j - \frac{1}{2})^{\beta_j} q_j^{\beta_j}\big) \end{aligned} \tag{51}$$

the approximation error can be expressed as

$$\begin{aligned} e_j(q_j, \epsilon_j) - \widehat{e}_j(q_j, \epsilon_j) = 2\epsilon_j \Big(&\sum_{i=2}^{+\infty} \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} \widetilde{f}_j(\xi) d\xi \\ &- \frac{\nu_j q_j^{p_j}}{2(p_j+1)} \Big(1 - Q_{1/\beta_j}\Big(\omega_j\big((\tau_j + \frac{1}{2})q_j\big)^{\beta_j}\Big)\Big)\Big). \end{aligned} \tag{52}$$

First, for every $i \geq 1$, we have

$$\begin{aligned} &\widetilde{f}_j\big((\tau_j + i - \frac{1}{2})q_j\big) \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} d\xi \\ &\quad \leq \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} \widetilde{f}_j(\xi) d\xi \\ &\quad \leq \widetilde{f}_j\big((\tau_j + i - \frac{3}{2})q_j\big) \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} d\xi \end{aligned} \tag{53}$$

$$\text{with} \quad \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} d\xi = \frac{\nu_j q_j^{p_j+1}}{p_j+1}. \tag{54}$$

In addition, we have the following inequalities:

$$\int_{(\tau_j + i - \frac{1}{2})q_j}^{(\tau_j + i + \frac{1}{2})q_j} \widetilde{f}_j(\xi) d\xi \quad \leq \quad q_j \widetilde{f}_j\big((\tau_j + i - \frac{1}{2})q_j\big) \tag{55}$$

and, for every $i \geq 2$,

$$q_j \widetilde{f}_j\big((\tau_j + i - \frac{3}{2})q_j\big) \leq \int_{(\tau_j + i - \frac{5}{2})q_j}^{(\tau_j + i - \frac{3}{2})q_j} \widetilde{f}_j(\xi) d\xi. \tag{56}$$

We deduce from (53), (54), (55), (56) and (51) that

$$\begin{aligned} &\frac{\nu_j q_j^{p_j}}{2(p_j+1)} \Big(1 - Q_{1/\beta_j}\big(\omega_j(\tau_j + \frac{1}{2})^{\beta_j} q_j^{\beta_j}\big)\Big) - \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j + \frac{1}{2})q_j\big) \\ &\quad \leq \sum_{i=2}^{+\infty} \int_{(\tau_j + i - \frac{3}{2})q_j}^{(\tau_j + i - \frac{1}{2})q_j} |\xi - r_{i,j}|^{p_j} \widetilde{f}_j(\xi) d\xi \\ &\quad \leq \frac{\nu_j q_j^{p_j}}{2(p_j+1)} \Big(1 - Q_{1/\beta_j}\big(\omega_j(\tau_j + \frac{1}{2})^{\beta_j} q_j^{\beta_j}\big)\Big) \\ &\qquad + \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j + \frac{1}{2})q_j\big). \end{aligned} \tag{57}$$

Therefore, the approximation error satisfies

$$\begin{aligned} -2\epsilon_j \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j + \frac{1}{2})q_j\big) &\leq e_j(q_j, \epsilon_j) - \widehat{e}_j(q_j, \epsilon_j) \\ &\leq 2\epsilon_j \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j + \frac{1}{2})q_j\big) \end{aligned} \tag{58}$$

which yields the desired approximation of the distortion.
Let us now focus on the expression of the distortion at high bitrate. When $q_j \to 0$, according to (48), the first term in the left hand side of (15) is such that

$$Q_{(p_j+1)/\beta_j}\big(\omega_j(\tau_j - \frac{1}{2})^{\beta_j} q_j^{\beta_j}\big) = O(q_j^{p_j+1}). \tag{59}$$

Moreover, using (53) and (54), we obtain

$$\begin{aligned} \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j + \frac{1}{2})q_j\big) &\leq \int_{(\tau_j - \frac{1}{2})q_j}^{(\tau_j + \frac{1}{2})q_j} |\xi - r_{1,j}|^{p_j} \widetilde{f}_j(\xi) d\xi \\ &\leq \frac{\nu_j q_j^{p_j+1}}{p_j+1} \widetilde{f}_j\big((\tau_j - \frac{1}{2})q_j\big) \end{aligned} \tag{60}$$

which shows that

$$\int_{(\tau_j - \frac{1}{2})q_j}^{(\tau_j + \frac{1}{2})q_j} |\xi - r_{1,j}|^{p_j} \widetilde{f}_j(\xi) d\xi = O(q_j^{p_j+1}). \tag{61}$$

In addition, we have

$$\begin{aligned} &\frac{\nu_j q_j^{p_j}}{2(p_j+1)} \Big(1 - Q_{1/\beta_j}\big(\omega_j(\tau_j + \frac{1}{2})^{\beta_j} q_j^{\beta_j}\big)\Big) \\ &\qquad = \frac{\nu_j q_j^{p_j}}{2(p_j+1)} \big(1 + O(q_j)\big). \end{aligned} \tag{62}$$

Since (16) shows that $e_j(q_j, \epsilon_j) - \widehat{e}_j(q_j, \epsilon_j) = O(q_j^{p_j+1})$, it can be deduced from (59), (61) and (62) that (17) holds.

Appendix C
Solution of the bit allocation problem

For simplicity, for every $j \in \{1, \ldots, J\}$, we will drop the index k in the variables a_j^k, c_j^k, γ_j^k, α_j^k, and δ_j^k, which are used in (13) and (18).

As $g_j(l_j)$ is a decreasing function of l_j for every $j \in \{1,\dots,J\}$, it is clear that, if $\sum_{j=1}^J \frac{n_j}{n} g_j(l_j^{b_j+1}) > R_{\max} \Leftrightarrow \sum_{j=1}^J \frac{n_j}{n}(a_j l_j^{b_j+1} + c_j) > R_{\max}$, then Problem $(\mathcal{P}_{\mathbf{b}})$ admits no solution since $C \cap ([l_1^{b_1}, l_1^{b_1+1}] \times \cdots \times [l_J^{b_J}, l_J^{b_J+1}])$ is empty. Another particular case is when

$$\sum_{j=1}^J \frac{n_j}{n}(a_j l_j^{b_j} + c_j) \leq R_{\max} \tag{63}$$

Since, for every $j \in \{1,\dots,J\}$, d_j is an increasing function, the solution to $(\mathcal{P}_{\mathbf{b}})$ is obviously $\widetilde{\mathbf{l}}_{\mathbf{b}} = (l_1^{b_1},\dots,l_J^{b_J})$.

In the following, we will discard these two trivial cases by assuming that

$$\sum_{j=1}^J \frac{n_j}{n}(a_j l_j^{b_j} + c_j) > R_{\max}$$
$$\text{and} \quad \sum_{j=1}^J \frac{n_j}{n}(a_j l_j^{b_j+1} + c_j) \leq R_{\max}. \tag{64}$$

Under these assumptions, since $(l_1^{b_1+1},\dots,l_J^{b_J+1}) \in C \cap ([l_1^{b_1}, l_1^{b_1+1}] \times \cdots [l_J^{b_J}, l_J^{b_J+1}])$, the intersection set is nonempty and the problem $(\mathcal{P}_{\mathbf{b}})$ has a solution $\widetilde{\mathbf{l}}_{\mathbf{b}}$. In order to find this solution, we will apply the Fenchel-Rockafellar duality theorem [44].

Theorem 1: Let f and g be two lower-semicontinuous convex functions from $\mathbb{R}^J$ to $]-\infty,+\infty]$. Then, provided that $\mathrm{dom}(f) \cap \mathrm{dom}(g)$ is nonempty, we have

$$\inf_{\mathbf{l}\in\mathbb{R}^J}(f(\mathbf{l}) + g(\mathbf{l})) = \max_{\mathbf{l}^*\in\mathbb{R}^J}(-g^*(-\mathbf{l}^*) - f^*(\mathbf{l}^*)), \tag{65}$$

where f^* (resp. g^*) is the convex conjugate of f (resp. g).[2] In our case, we take $g = \iota_C$ where ι_C is the indicator function[3] of the closed convex set C defined by (19). Taking $\mathbf{l} \in C$ is equivalent to take $\mathbf{l} \in \mathbb{R}^J$ such that

$$\mathbf{e}^\top \mathbf{l} \geq \sum_{j=1}^J \frac{n_j}{n} c_j - R_{\max}, \text{ with } \mathbf{e} = -\frac{1}{n}(n_1 a_1,\dots,n_J a_J) \tag{66}$$

Thus, the conjugate of g satisfies

$$\forall\, \mathbf{l}^* \in \mathbb{R}^J, \quad g^*(\mathbf{l}^*) = \sup_{\mathbf{l}\in C} \mathbf{l}^\top \mathbf{l}^* = \sup_{\mathbf{l}\in C}(\lambda \mathbf{l}^\top \mathbf{e} + \mathbf{l}^\top \mathbf{l}_\perp^*), \tag{67}$$

where $\mathbf{l}_\perp^*$ belongs to $\mathrm{Vect}\{\mathbf{e}\}^\perp$, the orthogonal subspace of $\mathbf{e}$, and $\lambda \in \mathbb{R}$. From (66), we see that if $\mathbf{l}_\perp^* \neq \mathbf{0}$, $g^*(\mathbf{l}^*) = +\infty$. Furthermore, if $\mathbf{l}^* = \lambda \mathbf{e}$ with $\lambda > 0$, the supremum over $\mathbf{l}$ of $\mathbf{l}^\top \mathbf{e}$ is infinite. Finally, we obtain for all $\mathbf{l}^* \in \mathbb{R}^J$

$$g^*(\mathbf{l}^*) = \begin{cases} \lambda\Big(\sum_{j=1}^J \frac{n_j}{n} c_j - R_{\max}\Big) & \text{if } \mathbf{l}^* = \lambda \mathbf{e} \text{ with } \lambda \leq 0 \\ +\infty & \text{else.} \end{cases}$$

On the other hand, we take, for every $\mathbf{l} \in \mathbb{R}^J$, $f(\mathbf{l}) = D(\mathbf{l}) + \iota_{\mathbf{P}_{\mathbf{b}}}(\mathbf{l})$, where $\mathbf{P}_{\mathbf{b}}$ is the box defined at the beginning of Section IV-B. Thus, f can be rewritten as

$$\forall\, \mathbf{l} \in \mathbb{R}^J, \quad f(\mathbf{l}) = \sum_{j=1}^J \phi_j(l_j) \tag{68}$$

where, for every $j \in \{1,\dots,J\}$,

$$\forall\, l_j \in \mathbb{R}, \;\; \phi_j(l_j) = \rho_j \epsilon_j(\alpha_j 2^{\gamma_j l_j} + \delta_j) + \iota_{[l_j^{b_j}, l_j^{b_j+1}]}(l_j). \tag{69}$$

Using the separability of the convex conjugate of f, we get

$$\forall\, \mathbf{l}^* = (l_1^*,\dots,l_J^*) \in \mathbb{R}^J \qquad f^*(\mathbf{l}^*) = \sum_{j=1}^J \phi_j^*(l_j^*). \tag{70}$$

For any given $j \in \{1,\dots,J\}$ and $l_j^* \in \mathbb{R}$, let us define

$$\forall\, l_j \in \mathbb{R}, \qquad \psi_j(l_j) = l_j l_j^* - \rho_j \epsilon_j(\alpha_j 2^{\gamma_j l_j} + \delta_j). \tag{71}$$

We can write

$$\phi_j^*(l_j^*) = \sup_{l_j^{b_j} \leq l_j \leq l_j^{b_j+1}} \psi_j(l_j). \tag{72}$$

Furthermore,

$$\forall\, l_j \in \mathbb{R},\; \psi_j'(l_j) = l_j^* - \ln(2)\rho_j \epsilon_j \gamma_j \alpha_j 2^{\gamma_j l_j} = l_j^* - \frac{\kappa_j N_j \gamma_j}{n} 2^{\gamma_j l_j}.$$

Thus, if $l_j^* \leq 0$, then $\psi_j'(l_j) < 0$ and $\phi_j^*(l_j^*) = \psi_j(l_j^{b_j})$. In turn, if $l_j^* > 0$, then it can be checked that $\psi_j'(l_j) < 0$ if and only if

$$l_j > \frac{1}{\gamma_j} \log_2\Big(\frac{n l_j^*}{\kappa_j N_j \gamma_j}\Big). \tag{73}$$

Three cases have then to be considered:

(i) If $l_j^{b_j} \geq \frac{1}{\gamma_j}\log_2\Big(\frac{n l_j^*}{\kappa_j N_j \gamma_j}\Big)$ which is equivalent to $2^{\gamma_j l_j^{b_j}} \geq \frac{n l_j^*}{\kappa_j N_j \gamma_j}$ then, for every $l_j \geq l_j^{b_j}$, $\psi_j'(l_j) < 0$ and

$$\phi_j^*(l_j^*) = \psi_j(l_j^{b_j}) = l_j^* l_j^{b_j} - \rho_j \epsilon_j(\alpha_j 2^{\gamma_j l_j^{b_j}} + \delta_j). \tag{74}$$

(ii) Similarly, if $\frac{n l_j^*}{\kappa_j N_j \gamma_j} \geq 2^{\gamma_j l_j^{b_j+1}}$ then, for every $l_j \in [l_j^{b_j}, l_j^{b_j+1}]$, $\psi_j'(l_j) > 0$ and

$$\phi_j^*(l_j^*) = \psi_j(l_j^{b_j+1}) = l_j^* l_j^{b_j+1} - \rho_j \epsilon_j(\alpha_j 2^{\gamma_j l_j^{b_j+1}} + \delta_j). \tag{75}$$

(iii) Otherwise, if $2^{\gamma_j l_j^{b_j}} < \frac{n l_j^*}{\kappa_j N_j \gamma_j} < 2^{\gamma_j l_j^{b_j+1}}$ then,

$$\begin{aligned} \phi_j^*(l_j^*) &= \psi_j\Big(\frac{1}{\gamma_j}\log_2\Big(\frac{n l_j^*}{\kappa_j N_j \gamma_j}\Big)\Big) \\ &= \frac{l_j^*}{\gamma_j \ln 2}\Big(\ln\Big(\frac{n l_j^*}{\kappa_j N_j \gamma_j}\Big) - 1\Big) - \rho_j \epsilon_j \delta_j. \end{aligned} \tag{76}$$

Now, by recalling that $\mathrm{dom}(g) = \{-\lambda \mathbf{e}, \lambda \geq 0\}$, the dual problem can be reexpressed as

$$\begin{aligned} \max_{\mathbf{l}^*\in\mathbb{R}^J}(-g^*(-\mathbf{l}^*) - f^*(\mathbf{l}^*)) &= \max_{-\mathbf{l}^*\in\mathrm{dom}(g)}(-g^*(-\mathbf{l}^*) - f^*(\mathbf{l}^*)) \\ &= \max_{\lambda\geq 0} \Phi(\lambda) \end{aligned} \tag{77}$$

where

$$\forall \lambda \in \mathbb{R}_+,\; \Phi(\lambda) = \lambda\Big(\sum_{j=1}^J \frac{n_j}{n} c_j - R_{\max}\Big) - \sum_{j=1}^J \phi_j^*\Big(-\lambda \frac{n_j}{n} a_j\Big).$$

[2]Recall that $\mathrm{dom}(f) = \{\mathbf{l} \in \mathbb{R}^J | f(\mathbf{l}) < +\infty\}$ and f^* is defined as: $\forall\, \mathbf{l}^* \in \mathbb{R}^J$, $f^*(\mathbf{l}^*) = \sup_{\mathbf{l}\in\mathbb{R}^J}(\mathbf{l}^\top \mathbf{l}^* - f(\mathbf{l}))$.

[3]The indicator function of C is defined as: $\forall\, x \in \mathbb{R}^J$, $\iota_C(x) = 0$ if $x \in C$; $+\infty$ otherwise.

According to (74)-(76) and the notation introduced in (22) and (23), Φ is the function defined in (30). The derivative of this function is given by

$$\forall \lambda \in \mathbb{R}_+, \qquad \Phi'(\lambda) = \sum_{j=1}^{J} \frac{n_j}{n} c_j - R_{\max} - \sum_{j=1}^{J} \varphi_j'(\lambda) \quad (78)$$

where φ_j' corresponds to the derivative of the function φ_j defined in Proposition 3. Thus, it can be checked that, for every $\lambda \in \mathbb{R}_+^*$, we have $\Phi''(\lambda) \leq 0$. The inequality being strict if and only if $\min_{1\leq j\leq J} \underline{\lambda}_j < \lambda < \max_{1\leq j\leq J} \overline{\lambda}_j$, Φ is strictly concave on this interval. In addition, if $\lambda \leq \min_{1\leq j\leq J} \underline{\lambda}_j$, then

$$\Phi'(\lambda) = \sum_{j=1}^{J} \frac{n_j}{n} (a_j l_j^{b_j} + c_j) - R_{\max} > 0 \quad (79)$$

and, if $\lambda \geq \max_{1\leq j\leq J} \overline{\lambda}_j$, then

$$\Phi'(\lambda) = \sum_{j=1}^{J} \frac{n_j}{n} (a_j l_j^{b_j+1} + c_j) - R_{\max} \leq 0 \quad (80)$$

where the assumptions given by (64) have been used.
As Φ' is strictly decreasing on $[\min_{1\leq j\leq J} \underline{\lambda}_j, \max_{1\leq j\leq J} \overline{\lambda}_j]$, we deduce that there exists a unique value $\widetilde{\lambda}$ in this interval such that $\Phi'(\widetilde{\lambda}) = 0$. Thus, $\widetilde{\lambda}$ corresponds to the maximizer of Φ over $\mathbb{R}_+$. From the definitions of the sets in (27), (28) and (29), we get:

$$\forall j \in \mathbb{I}, \qquad \widetilde{\lambda} \leq \underline{\lambda}_j \quad (81)$$

$$\forall j \in \mathbb{J}, \qquad \underline{\lambda}_j < \widetilde{\lambda} \leq \overline{\lambda}_j \quad (82)$$

$$\forall j \in \mathbb{K}, \qquad \widetilde{\lambda} > \overline{\lambda}_j. \quad (83)$$

Finally, it can be deduced from (78) that $\Phi'(\widetilde{\lambda}) = 0$ implies that

$$\sum_{j\in\mathbb{J}} N_j \log_2\left(\frac{\widetilde{\lambda}}{\kappa_j}\right) + \sum_{j\in\mathbb{I}} N_j \gamma_j l_j^{b_j} + \sum_{j\in\mathbb{K}} N_j \gamma_j l_j^{b_j+1} = \sum_{j=1}^{J} n_j c_j - n R_{\max} \quad (84)$$

which yields the expression of $\widetilde{\lambda}$ in (25).

Furthermore, the optimal value $\widetilde{\mathbf{l}}_{\mathbf{b}} = (\widetilde{l}_{1,\mathbf{b}}, \ldots, \widetilde{l}_{J,\mathbf{b}})$ of $\mathbf{l}$ is given by the critical point of f. This means that, for every $j \in \{1, \ldots, J\}$, $\widetilde{l}_{j,\mathbf{b}}$ is the maximizer of ψ_j over $[l_j^{b_j}, l_j^{b_j+1}]$ when $l_j^* = -\widetilde{\lambda} n_j a_j / n$. Therefore, we get the optimal values $\widetilde{l}_{j,\mathbf{b}}$ given by (24).

References

[1] M. Kaaniche, A. Fraysse, B. Pesquet-Popescu, and J.-C. Pesquet, "A convex programming bit allocation method for sparse sources," in *Picture Coding Symposium*, Poland, May 2012, pp. 277–280.

[2] P. Westerink, J. Biemond, and D. Boekee, "An optimal bit allocation algorithm for sub-band coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York, USA, 1988, pp. 757–760.

[3] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[4] Y. Shoham and A. Gersho, "Efficient codebook allocation for an arbitrary set of vector quantizers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, Florida, USA, 1985, pp. 1696–1699.

[5] S.-W. Wu and A. Gersho, "Rate-constrained optimal block-adaptive coding for digital tape recording of HDTV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, no. 1, pp. 100–112, 1991.

[6] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 533–545, 1994.

[7] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion based video compression: optimal video frame compression and object boundary encoding*. Kluwer Academic Publishers, 1997.

[8] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis-based buffered compression and fast approximations," *IEEE Transactions on Image Processing*, vol. 3, no. 1, pp. 26–40, 1994.

[9] E. A. Riskin, "Optimal bit allocation via the generalized BFOS algorithm," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 400–402, 1991.

[10] K. Ferguson and N. Allinson, "Modified steepest-descent for bit allocation in strongly dependent video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 7, pp. 1057–1062, 2009.

[11] Y. Sermadevi and S. S. Hemami, "Convexity results for a predictive video coder," in *Asilomar Conference on Signal, Systems and computers*, vol. 2, CA, November 2004, pp. 1713–1717.

[12] T. André, M. Cagnazzo, M. Antonini, and M. Barlaud, "JPEG2000-compatible scalable scheme for wavelet-based video coding," *EURASIP Journal on Image and Video Processing*, vol. 2007, p. 11, 2007.

[13] P. Noll and R. Zelinski, "Bounds on quantizer performance in the low bit-rate region," *IEEE Trans. on Communications*, vol. 26, no. 2, pp. 300–304, 1978.

[14] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-Gaussian memoryless sources," *IEEE Transactions on Information Theory*, vol. 30, no. 3, pp. 485–497, 1984.

[15] A. György, T. Linder, and K. Zeger, "On the rate-distortion function of random vectors and stationary sources with mixed distributions," *IEEE Trans. on Information Theory*, vol. 45, no. 6, pp. 2110–2115, 1999.

[16] T. André, M. Antonini, M. Barlaud, and R. Gray, "Entropy-based distortion measure and bit allocation for wavelet image compression," *IEEE Trans. on Image Processing*, vol. 16, no. 12, pp. 3058–3064, 2007.

[17] G. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. on Information Theory*, vol. 42, no. 5, pp. 1365–1374, 1996.

[18] D. Marco and D. Neuhoff, "Low-resolution scalar quantization for Gaussian sources and squared error," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1689–1697, 2006.

[19] J. Sun, W. Gao, D. Zhao, and Q. Huang, "Statistical model, analysis and approximation of rate-distortion function in MPEG-4 FGS videos," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 4, pp. 535–539, 2006.

[20] M. Wang and M. Van Der Schaar, "Operational rate-distortion modeling for wavelet video coders," *IEEE Trans. on Image Processing*, vol. 54, no. 9, pp. 3505–3517, 2006.

[21] M. Gaubatz and S. Hemami, "Efficient entropy estimation based on doubly stochastic models for quantized wavelet image data," *IEEE Trans. on Image Processing*, vol. 16, no. 4, pp. 967–981, 2007.

[22] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 676–683, 1968.

[23] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, November 1998.

[24] G. J. Peng and W. L. Whang, "Inter-layer bit-allocation for scalable video coding," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2592–2606, May 2012.

[25] F. Abramovitch, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a bayesian approach," *Journal of the Royal Statistical Society*, vol. 60, pp. 725–749, 1998.

[26] A. Antoniadis, D. Leporini, and J.-C. Pesquet, "Wavelet thresholding for some classes of non-Gaussian noise," *Statistica Neerlandica*, vol. 56, no. 4, pp. 434–453, 2002.

[27] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[28] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. on Information Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[29] A. Fraysse, B. Pesquet Popescu, and J.-C. Pesquet, "On the uniform quantization of a class of sparse sources," *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 3243–3263, 2009.

[30] W. Szepanski, "Δ-entropy and rate distortion bounds for generalized Gaussian information sources and their application to image signals," *Electronics Letters*, vol. 16, no. 3, pp. 109–111, 1980.

[31] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.

[32] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1158–1170, 2000.

[33] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[34] M. Rabbani and R. R. Joshi, "An overview of the JPEG 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.

[35] B. Usevitch, "Optimal bit allocation for biorthogonal wavelet coding," in *Data Compression Conference*, Snowbird, USA, March 1996, pp. 387–395.

[36] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*. Academic Press, 2000, pp. 669–684.

[37] W. Gautschi, "The incomplete Gamma functions since Tricomi," in *Tricomi's ideas and contemporary applied mathematics, Atti dei Convegni Lincei*, no. 147, Accademia Nazionale dei Lincei, Roma, 1998, pp. 203–237.

[38] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu, "Distortion evaluation in transform domain for adaptive lifting schemes," in *International Workshop on Multimedia Signal Processing*, Cairns, Queensland, Australia, October 2008.

[39] I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of methods of applied statistics*. New York: Wiley, 1967.

[40] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," ITU SG16 VCEG-M33, Austin, TX, USA, Tech. Rep., April 2001.

[41] C. Tian and S. S. Hemami, "An embedded image coding system based on tarp filter with classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Montreal, Canada, 2004, pp. 49–52.

[42] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[43] L. S. Gradshteyn and L. M. Ryzhik, *Tables of Integrals, Series and Products*. San Diego: Academic press, 2000.

[44] R. Rockafellar, *Convex analysis*. Princeton Landmarks in Mathematics, 1997.

TABLE I

THE AVERAGE PSNR DIFFERENCES AND THE BITRATE SAVING AT LOW AND HIGH BITRATES. THE GAIN OF THE PROPOSED APPROACH W.R.T THE STATE-OF-THE-ART METHOD [12].

	bitrate saving (in %)		PSNR gain (in dB)	
Images	low	high	low	high
goldhill	-4.89	-14.50	0.16	0.82
barbara	-31.94	-12.80	1.34	0.96
lena	-1.64	-8.79	0.01	0.45
einst	-8.20	-19.42	0.15	0.65
marseille	-6.96	-13.86	0.24	0.68
straw	-16.41	-19.53	0.51	0.77
elaine	-0.45	-16.37	0.02	0.69
castle	-7.78	-9.02	0.40	0.68
cartoon	-5.16	-4.27	0.36	0.63
average	-9.27	-13.17	0.35	0.70

TABLE II

PERFORMANCE COMPARISON IN TERMS OF PSNR AFTER ENTROPY ENCODING PROCESS.

Image	Rate (bpp)	JPEG2000	Our method with AC-1	Our method with AC-2
goldhill	0.15	**28.93**	28.14	28.36
	0.25	**30.44**	29.79	29.94
	0.5	**33.15**	32.11	32.36
lena	0.15	**31.51**	30.87	31.13
	0.25	**33.98**	32.95	33.26
	0.5	**37.16**	35.88	36.16
straw	0.15	20.47	20.68	**20.83**
	0.25	21.89	21.87	**21.99**
	0.5	24.35	24.25	**24.40**
marseille	0.15	24.67	24.82	**24.89**
	0.25	26.34	26.31	**26.39**
	0.5	29.02	28.97	**29.06**
einst	0.15	**30.22**	29.55	29.71
	0.25	**31.32**	30.77	30.91
	0.5	**32.92**	32.25	32.46

Mounir Kaaniche received the Engineering and Master degrees in telecommunications from the SUP'COM Tunis, Tunisia, in 2006 and 2007, and the Ph.D. degree in signal and image processing from Télécom ParisTech, Paris, France, in 2010.

In 2011, he was a post-doctoral researcher at Télécom ParisTech. From Oct. 2012 to Aug. 2013, he worked as a Graduate Teaching Assistant with Université Paris-Est Marne-la-Vallée, France. Since Sept. 2013, he is an assistant professor at Université Paris 13, Villetaneuse, France. His current research interests are in still/stereo image compression, adaptive wavelets and image retrieval.

Aurélia Fraysse was born in Creteil, France, in 1978. She graduated from University Paris 12 in 2001. She received her Ph. D. degree in mathematics at University Paris 12 in 2005.

In 2006, she was a research assistant at Ecole Nationale Supérieure des Télécommunications (Télécom Paris Tech). She is presently associate professor at IUT Cachan and researcher with the Laboratoire des Signaux et Systèmes (CNRS-Supelec-Univ. Paris-Sud).

Béatrice Pesquet Popescu (SM'06, F'13) received the engineering degree in telecommunications from the "Politehnica" Institute in Bucharest in 1995 (highest honours) and the Ph.D. degree from the Ecole Normale Supérieure de Cachan in 1998.

In 1998, she was a Research and Teaching Assistant with Université Paris XI, Paris. In 1999, she joined Philips Research France, Suresnes, France, where she worked for two years as a Research Scientist, then as a Project Leader, in scalable video coding. Since Oct. 2000 she is with Télécom ParisTech (formerly, ENST), first as an Associate Professor, and since 2007 as a Professor, Head of the Multimedia Group. She is the Head of the UBIMEDIA common research laboratory between Alcatel-Lucent and Institut Télécom. Her current research interests are in source coding, scalable, robust and distributed video compression and sparse representations.

Dr. Pesquet-Popescu was an EURASIP BoG member (2003-2010), and is an IEEE Signal Processing Society IVMSP TC member and MMSP TC associate member. She serves in the Editorial Board of IEEE Signal Processing Magazine, and as an Associate Editor for IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on CSVT, Elsevier Image Communication, and Hindawi Int. J. Digital Multimedia Broadcasting journals and was till 2010 an Associate Editor for Elsevier Signal Processing. She was a Technical Co-Chair for the PCS2004 conference, and General Co-Chair for IEEE SPS MMSP2010, EUSIPCO 2012, and IEEE SPS ICIP 2014 conferences. Beatrice Pesquet-Popescu is a recipient of the "Best Student Paper Award" in the IEEE Signal Processing Workshop on Higher-Order Statistics in 1997, of the Bronze Inventor Medal from Philips Research and in 1998 she received a "Young Investigator Award" granted by the French Physical Society. She holds 25 patents in video coding and has authored more than 280 book chapters, journal and conference papers in the field. In 2006, she was the recipient, together with D. Turaga and M. van der Schaar, of the IEEE Trans. on Circuits and Systems for Video Technology "Best Paper Award" and in 2012 Best Associate Editor of the same journal.

Jean-Christophe Pesquet (S'89–M'91–SM'99–F'12) received the engineering degree from Supélec, Gif-sur-Yvette, France, in 1987, the Ph.D. degree from University Paris-Sud (XI), Paris, France, in 1990, and the Habilitation à Diriger des Recherches from University Paris-Sud in 1999.

From 1991 to 1999, he was a Maître de Conférences at University Paris-Sud, and a Research Scientist at the Laboratoire des Signaux et Systèmes, Centre National de la Recherche Scientifique (CNRS), Gif sur Yvette. He is currently a Professor (classe exceptionnelle) with Université Paris-Est Marne-la-Vallée, France and the Deputy Director of the Laboratoire d'Informatique of the university (UMR–CNRS 8049).

Fig. 3. Test images used for the experiments.

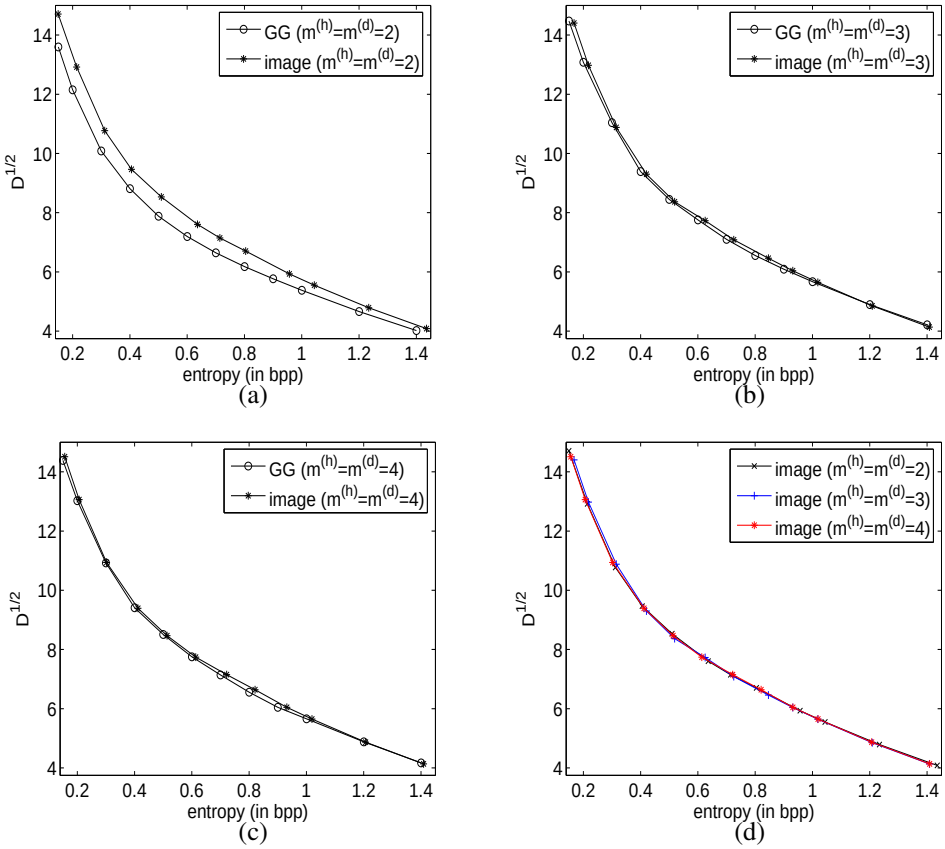


Fig. 4. $D^{1/2}$ versus entropy (in bpp) for a uniform scalar quantizer with a deadzone of size q_j (i.e. $\tau_j = 1$) for "marseille" image: influence of the number of intervals.

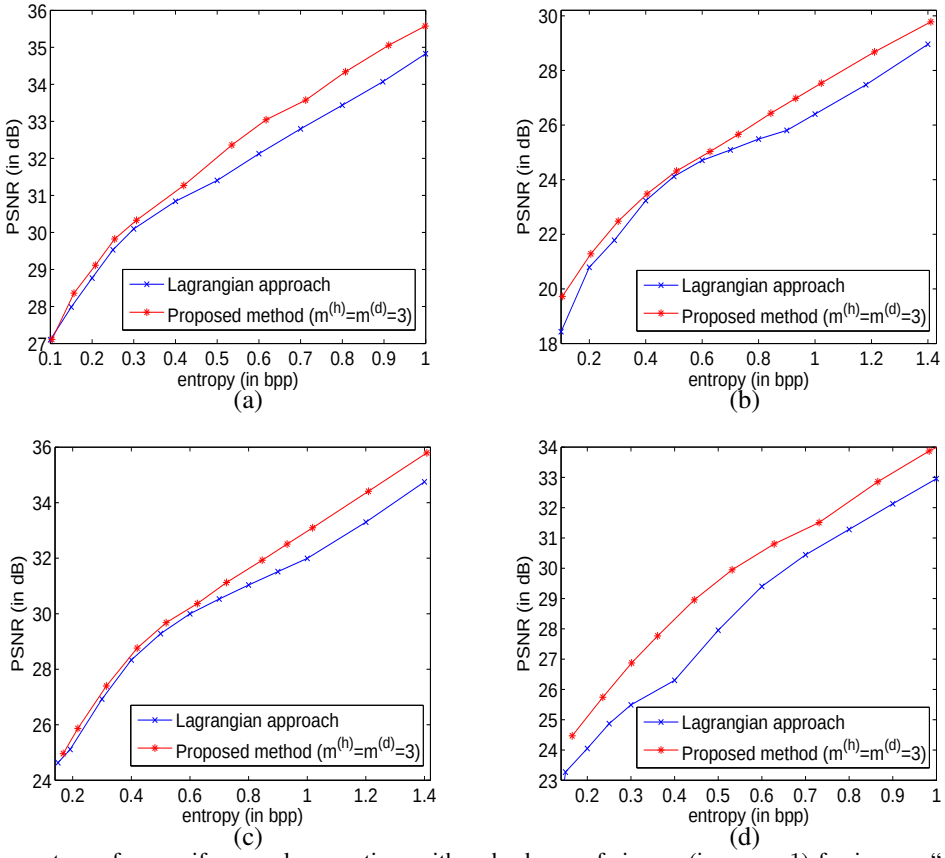


Fig. 5. PSNR versus entropy for a uniform scalar quantizer with a deadzone of size q_j (i.e. $\tau_j = 1$) for images "goldhill" (a), "straw" (b), "marseille" (c) and "barbara" (d): performance of the proposed approach vs the Lagrangian one.

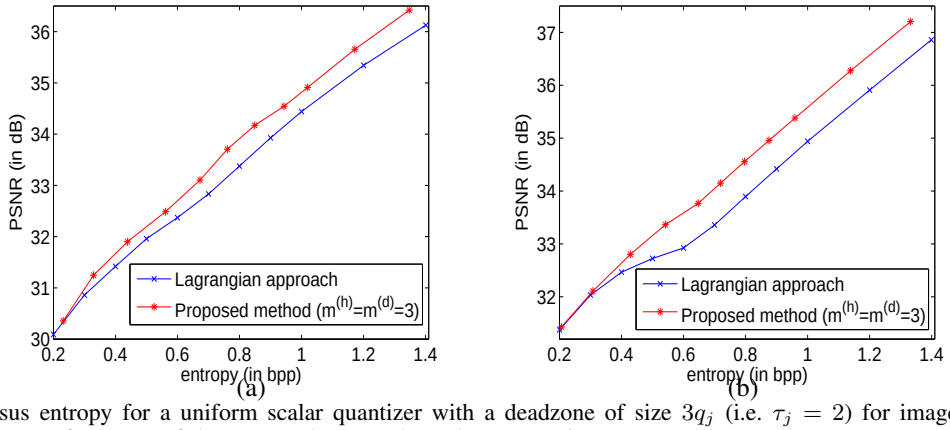
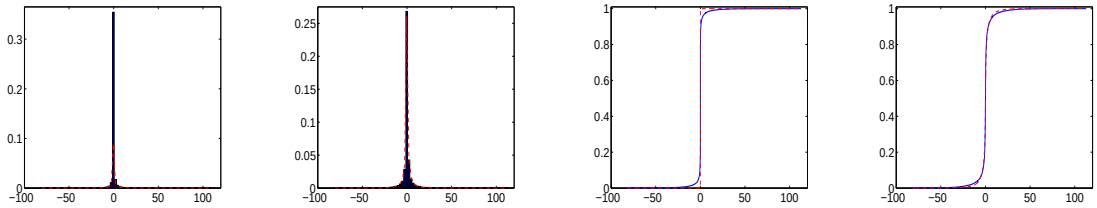


Fig. 6. PSNR versus entropy for a uniform scalar quantizer with a deadzone of size $3q_j$ (i.e. $\tau_j = 2$) for images "einst" (left side) and "elaine" (right side): performance of the proposed approach vs the Lagrangian one.



(a) $\beta_j = 0.05$ (b) $\beta_j = 0.34$, $\epsilon_j = 0.40$ (c) KS=0.19 (d) KS=0.06

Fig. 7. Modelling the distribution of the diagonal detail wavelet coefficient of the "cartoon" image using (a) GG model (b) BGG model. The cumulative distribution function using (c) GG model (d) BGG model. The curve plotted in solid (resp. dashed) line is associated with the subband wavelet coefficients (resp. theoretical model).

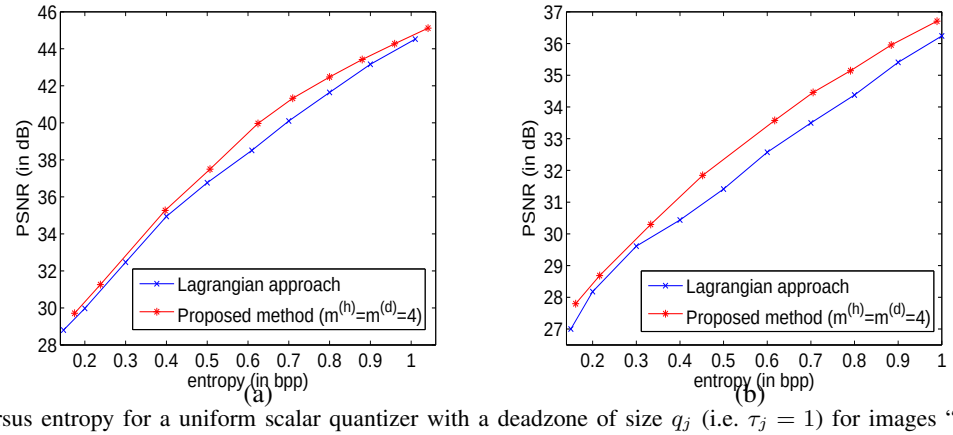


Fig. 8. PSNR versus entropy for a uniform scalar quantizer with a deadzone of size q_j (i.e. $\tau_j = 1$) for images "cartoon" (a) and "castle" (b): performance of the proposed approach vs the Lagrangian one.

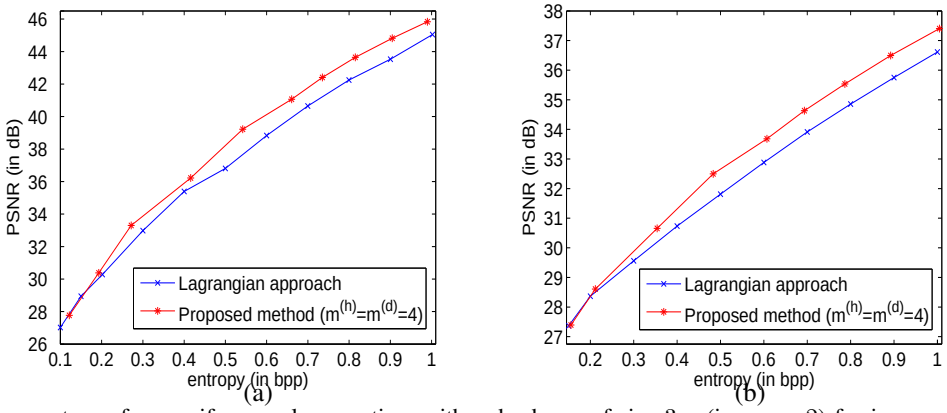


Fig. 9. PSNR versus entropy for a uniform scalar quantizer with a deadzone of size $3q_j$ (i.e. $\tau_j = 2$) for images "cartoon" (a) and "castle" (b): performance of the proposed approach vs the Lagrangian one.

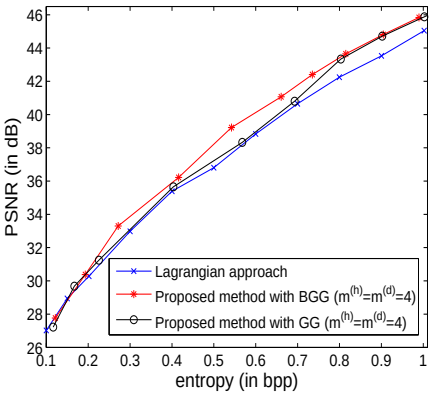


Fig. 10. PSNR versus entropy for a uniform scalar quantizer with a deadzone of size $3q_j$ (i.e. $\tau_j = 2$) for "cartoon" image: BGG model versus the GG one.