In-loop atom modulus quantization for matching pursuit and its application to video coding

Christophe De Vleeschouwer

Laboratoire de Télécommunications

Université catholique de Louvain, Belgium

Avideh Zakhor

Department of Electrical Engineering & Computer Sciences

University of California Berkeley, USA

E-mail: devlees@tele.ucl.ac.be, avz@eecs.berkeley.edu, Fax: +32 10 47 20 89

Abstract

This paper provides a precise analytical study of the selection and modulus quantization of matching pursuit (MP) coefficients. We demonstrate that an optimal rate-distortion trade-off is achieved by selecting the atoms up to a quality-dependent threshold, and by defining the modulus quantizer in terms of that threshold. In doing so, we take into account quantization error re-injection resulting from inserting the modulus quantizer inside the MP atom computation loop. *In-loop* quantization not only improves coding performance, but also affects the optimal quantizer design for both uniform and non-uniform quantization. We measure the impact of our work in the context of video coding. For both uniform and non-uniform quantization, the precise understanding of the relation between atom selection and quantization results in significant improvements in terms of coding efficiency. At high bitrates, the proposed non-uniform quantization scheme results in 0.5 to 2 dBs improvement over [8].

I. Introduction

Matching pursuit (MP) is a greedy and iterative approximation algorithm that generates a sparse representation of a signal with respect to an overcomplete set of basis functions [6]. The MP expansion is defined in terms of index, sign, and modulus of a subset of basis functions. The transmission of this representation requires both quantization and coding of the index and

Dr. C. De Vleeschouwer was funded by the Belgian NFS.

Prof. A. Zakhor was funded by NSF Grants CCR9903368, and ANI-9905799.

modulus information. As the set of basis functions is finite, the index is generally not quantized before entropy coding. On the contrary, the modulus information has to be quantized. In this paper we primarily deal with the entropy constrained modulus quantization of matching pursuit coefficients.

Our study relies on two main assumptions. Firstly, linear reconstruction is assumed. It is known that quantized overcomplete expansions could benefit from nonlinear iterative reconstruction [3],[5]; in practice however, complexity considerations dictate linear reconstruction in most applications. Secondly, independent scalar moduli quantization is considered. This is applicable to scenarios where successively transmitted atom modulus are independent, i.e. where the atom coding order is not driven by the atom modulus but rather by the entropy gain that can be achieved on the atom indices. This is for example the case when MP is used to encode motion residual images, also called displaced frame differences (DFD), in a hybrid video coder [1]. In this context, significant entropy gains are obtained by appropriate differential description of the atom positions on the DFD. This has motivated our study and differentiates it from [2], which is dedicated to systems coding the atoms in decreasing order of magnitude.

A previous work has already addressed the MP atom modulus quantization issue in the context of linear reconstruction and modulus-independent atom coding order [8]. This work has shown empirically that the optimal uniform quantizer stepsize for a MP expansion is proportional to the expansion dead-zone threshold, i.e. to the smallest modulus value of the expansion. Understanding this empirical result has been our initial motivation. Ultimately, besides providing an analytical derivation of the empirical results observed in [8], our work also refines and completes them. An original aspect of our study is to take into account the quantization error re-injection resulting from inserting the modulus quantizer inside the MP atom computation loop. It is known that, due to re-injection, the quantization error of an atom is likely to be corrected by subsequent iterations, resulting in improved coding efficiency [3]. Our work demonstrates that re-injection impacts the quantizer design. Specifically, re-injection affects the stepsize of the uniform quantizer, and makes the optimal entropy constrained quantizer non-uniform. We discuss how to design both the optimal uniform and non-uniform quantizers, and show that the non-uniform quantizer outperforms the uniform one at high bitrates. An additional contribution of our paper lies in the accurate statement of an atom selection stopping criterion. To ensure R-D optimality, atoms should be selected up to a quality-dependent threshold. In particular, we show that, when the signal to be expanded is partitioned into subspaces, the stopping criterion has to be verified in every subspace.

As justified along the text, our study relies on two main observations. Firstly, it uses the fact that atoms are selected mostly in decreasing order of magnitude. Secondly, to simplify the quantizer design, it makes the high-resolution assumption, which means that first order developments hold, and it assumes a near-constant expected atom coding cost.

The rest of the paper is organized as follows. Sections II and III express rate-distortion constraints, respectively for the expansion process and for the quantization stage of the MP coder. Section II shows that an optimal MP expansion must stop as soon as the moduli of the extracted atoms become smaller than a threshold. We refer to this rule as to the stopping criterion. In Section III, we assume an exponential distribution of the MP atom modulus, and review rate-constrained optimal quantization of such a random variable. We then consider the re-injection of the quantization error into the MP expansion loop, and show that re-injection reduces the impact of quantization error on the final reconstructed signal distortion. A formal analysis of that phenomenon reveals that the benefit of the re-injection is stronger for the atoms that are selected during the earlier MP iterations. As these atoms are also expected to have the largest moduli, we propose the use of a non-uniform quantizer and present an optimal design methodology. Section IV combines the results obtained in Sections II and III to derive a relation between the atom modulus expansion threshold and the quantization stepsize. The relation is simplified and discussed in the context of a hybrid video codec. Section V measures the impact of our work on video coding, and Section VI concludes. For clarity purposes, Tables I and II summarize the symbols and acronyms defined and used throughout the paper.

Symbol	Definition
f	MP source vector
$\mathcal{R}_n f$	Residual signal after n MP iterations
φ_{k_n}	Dictionary function selected at step n, with $n \geq 0$
α_n	Atom modulus at step n, i.e. $\alpha_n = \mid < \varphi_{k_n}, \mathcal{R}_n f > \mid$.
γ_n	Fraction of the residual energy extracted at step n, i.e. $\gamma_n = \alpha_n^2 / \|\mathcal{R}_n f\|^2$.
$\hat{\alpha}_n$	Quantized atom modulus at step n.
σ_n^2	Atom modulus quantization distortion at step n.
$\xi_{Q(.)}()$	Quantization error for a quantizer defined by the mapping function $Q(.)$.
Ω	Uniform quantizer stepsize.
δ	Quantizer reconstruction offset.
Ω_i	i^{th} bin quantizer stepsize.
λ	Lagrangian multiplier.
D	Distortion.
R	Rate.
R_{last}	Expected rate for the last selected atom.
f_X	pdf of random variable X.
D_X	Quantization distortion of the random variable X.
H_X	Entropy of the quantized random variable X.
Θ	Atom selection stopping threshold
M_Θ	Θ-subtracted modulus random variable.
μ	Mean value of the exponential Θ-subtracted modulus distribution.
E_θ	Residual energy after all atoms larger than θ have been selected
ρ	Residual energy decrease parameter, i.e. E_θ decreases as the power of ρ of θ.
θ_i	Lower boundary of the i^{th} quantizer bin.
$\beta'(\Theta)$	Uniform quantizer correction factor due to quantization error re-injection, see (21).
$\beta'_i(\Theta)$	Correction factor in (30) due to quantization error re-injection for the i^{th} bin of the R-D optimal non-uniform quantizer.
χ	Sequence of factors $\{\chi_i\}_{i>0}$ defined as the numerical solution to (54).
τ	Truncated version of the χ-sequence defined by (50) so that $\tau_i \approx \sqrt{{\beta'}_i^{-1}(\Theta)}$, i.e. such that $\Omega_i/\Theta \approx 0.66.\tau_i$ for the i^{th} bin of the R-D optimal quantizer.

TABLE I

SYMBOLS SUMMARY DEFINITION.

Acronym	Definition
MP	Matching Pursuit
DFD	Displaced Frame Difference
ECSQ	Entropy Constrained Scalar Quantization
R/D	Rate/Distortion
ISE	Iterative β'-Sequence Estimation (see Section IV-B.1)
RSC	Recursive χ-Sequence Computation (see Section IV-B.1)
RC-RSC	Reduced Complexity RSC (see Section IV-B.2)
UQ	Uniform Quantizer
UQ,IS	UQ with Immediate Stop of the search (see Section V)
ULQ	Uniform in-Loop Quantizer
NULQ	Non-Uniform in-Loop Quantizer

TABLE II
ACRONYMS.

II. STOPPING THRESHOLD FOR RATE/DISTORTION OPTIMAL ATOM SELECTION

Matching pursuit (MP) is a greedy and iterative expansion process. In a coding context, a critical question is when to stop the process. This section shows that an expansion is optimal in the rate/distortion (R/D) sense if it strictly captures all atoms larger than a threshold. As atoms are selected mostly in decreasing order of magnitude, this means that optimality is achieved by selecting atoms until their moduli become smaller than a threshold. So, in concrete terms, this section defines a stopping criterion. The stopping threshold depends on both the quantization and relative importance of distortion and rate. For practical purposes, it is important to note that when the signal to expand is partitioned into smaller subspaces, i.e. into blocks of pixels, for complexity reasons [7], the stopping criterion has to be met in every subspace.

Let a dictionary $\mathcal{D} = \{\varphi_k\}_{k=1}^{S} \in R^M$ be a frame such that $||\varphi_k|| = 1$ for all k. Given a source vector $f \in R^M$, MP is a greedy and iterative algorithm that approximates f by a linear combination of elements of $\mathcal{D}$ [6]. At step i, the algorithm selects the dictionary function φ_{k_i} that maximizes $|< \varphi_{k_i}, \mathcal{R}_i f >|$ and generates the residue for the next iteration, i.e. $\mathcal{R}_{i+1} f = \mathcal{R}_i f - < \varphi_{k_i}, \mathcal{R}_i f > \varphi_{k_i}$. In the initial step, $\mathcal{R}_0 f = f$. After n steps, the approximate signal is $f \approx \sum_{i=0}^{n-1} (s_i \,.\, \alpha_i \,.\, \varphi_{k_i})$, where $\alpha_i = | < \varphi_{k_i}, \mathcal{R}_i f > |$ and $s_i = sign(< \varphi_{k_i}, \mathcal{R}_i f >)$. The output of the MP expansion does not only contain the coefficients moduli $(\alpha_0, \alpha_1, \ldots)$, but also their signs and indices $(k_0, k_1, \ldots)$. An element of the decomposition, i.e. $< \varphi_{k_i}, \mathcal{R}_i f > \varphi_{k_i}$, is usually referred to as an atom, and $| < \varphi_{k_i}, \mathcal{R}_i f > |$ as its modulus.

In a coding context, it is desirable to achieve the highest quality at the lowest cost. A convenient way to formalize this problem is to use a Lagrange multiplier λ to define the relative importance of the distortion D and of the number of bits R. Given λ, the optimal rate-distortion trade-off is then the one that minimizes the Lagrangian cost function $\mathcal{L}(\lambda) = D + \lambda R$.

For an MP expansion, the atoms are selected mostly in decreasing order of magnitude. So, letting D_i denote the residual energy after i atoms have been selected, the distortion change $\Delta D_i = D_i - D_{i-1}$ due to the i^{th} atom is negative, but mostly increases as a function of the iteration index i. Moreover, an MP expansion is sparse, so that the number of bits devoted to an additional atom is more or less independent of the iteration index. Letting R_i denote the rate for i atoms, the rate increment $\Delta R_i = R_i - R_{i-1}$ due to the i^{th} atom is more or less constant as a function of i. As a consequence, the incremental contribution $\Delta D_i + \lambda \Delta R_i$ of successive atoms to $\mathcal{L}(\lambda)$ can be assumed to be mostly increasing as a function of i. This statement is true most of the time, and we assume that it ensures global convexity of the (R, D) curve drawn along the expansion, i.e. the curve that joins the (R_i, D_i) points resulting from the selection of i atoms. This also makes the selection of an additional atom worthwhile only until the j^{th} iteration, with

$$\Delta D_j + \lambda \Delta R_j = 0, \tag{1}$$

defining a stopping criterion for the MP expansion in terms of incremental rate and MP residue energy of the j^{th} atom.

To better understand the criterion, we express ΔD_j and ΔR_j for the j^{th} additional atom. By definition of the MP expansion procedure, any additional atom is orthogonal to the residual signal. Hence, letting α_j denote the modulus of the additional atom, and defining $\xi_{Q(.)}(\alpha_j)$ to be its quantization error for a quantization method defined by $Q(.)$, the decrease ΔD_j of the MP residue energy is defined in terms of the atom energy α_j^2 and squared quantization error $\xi^2_{Q(.)}(\alpha_j)$, i.e.

$$\Delta D_j = -\left(\alpha_j^2 - \xi^2_{Q(.)}(\alpha_j)\right) \tag{2}$$

ΔR_j can be divided in two parts: the first one $R_{Q(.)}(\alpha_j)$ corresponds to the quantized atom modulus; the second one ΔR_{index} corresponds to the rate for the sign and index of the additional atom, and is not directly affected by the quantization method. Based on Equations (1) and (2), the stopping criterion can be formulated in terms of the Lagrange multiplier λ, and the threshold modulus $\Theta \triangleq \alpha_j$ beyond which it is worthless to select an additional atom,

$$\lambda = \frac{-\Delta D_j}{\Delta R_j} = \frac{\Theta^2 - \xi^2_{Q(.)}(\Theta)}{\Delta R_{index} + R_{Q(.)}(\Theta)} \tag{3}$$

where $\xi_{Q(.)}(\Theta)$ is the quantization error of modulus Θ. The stopping criterion states that, to achieve a rate constrained optimal representation, atoms have to be selected until there are no atoms with moduli larger than Θ on any part of the residue. Note that, for a given λ, Θ depends on the quantizer $Q(.)$. A condition for overall optimality is to design the quantizer so that it achieves R/D optimality for the same Lagrange multiplier λ. The quantizer design is investigated in the next section.

III. R/D OPTIMAL INDEPENDENT ATOM MODULUS QUANTIZATION

In this section, we consider the quantization of a set of atoms in the R/D framework. At constant number of atoms, a coarser modulus quantization increases the distortion D but decreases the bit budget R of the MP expansion. The goal is to find the quantizer that minimizes the Lagrangian cost function $\mathcal{L}(\lambda) = D + \lambda R$ for a given multiplier λ, i.e. for a given relative importance of distortion and entropy.

We consider independent scalar quantization of the atoms. This is suited to a coding framework that deals with the atoms in a random order of modulus. As a consequence, the atom moduli can be viewed as independent samples of a random variable. This assumption is motivated by the fact that efficient entropy coding of the atom indices, i.e. position and shape, generally dictates their coding order [1]. Section V validates and further discusses this assertion.

This section proceeds as follows. First we define the notion of entropy-constrained, or R/D optimal, quantization of a random variable. Then, as the dead-zone subtracted atom modulus distribution closely fits an exponential model, we review the characteristics of the R/D optimal quantizer of an exponential random variable. Finally, we study how the re-injection of the quantization error into the MP expansion loop affects the optimal quantizer design. This is a major contribution of our paper and shows that the optimal quantizer is non-uniform.

A. Entropy-constrained scalar quantization: a definition

Consider an input random variable X having a smooth probability distribution function (pdf) $f_X(x)$ which is greater than zero over $(0, \infty)$ and zero elsewhere. The scalar quantizer $Q_X(.)$ is characterized in terms of a sequence of steps sizes $\{\Omega_i\}_{i \geq 0}$ and a sequence of reconstruction offsets $\{\delta_i\}_{i \geq 0}$. It is defined as a functional mapping of an input value $x > 0$ onto its output approximation $Q_X(x)$ such that

$$Q_X(x) = \sum_{i=0}^{j-1} \Omega_i + \delta_j, \text{ for } \sum_{\mathrm{i=0}}^{\mathrm{j-1}} \Omega_{\mathrm{i}} \leq \mathrm{x} < \sum_{\mathrm{i=0}}^{\mathrm{j}} \Omega_{\mathrm{i}} \tag{4}$$

The probability of the output value $y_j = \sum_{i=0}^{j-1} \Omega_i + \delta_j$ is

$$p_j = \int_{\sum_{i=0}^{j-1} \Omega_i}^{\sum_{i=0}^{j} \Omega_i} f_X(x) dx \tag{5}$$

The squared-error distortion of the quantizer is

$$D_X = E_X\{(X - Q_X(x))^2\} = \int_0^\infty (x - Q_X(x))^2 f(x) dx \tag{6}$$

The output entropy of the quantizer is

$$H_X = -\sum_{j \geq 0} p_j \log_2 p_j \tag{7}$$

Using a Lagrange multiplier λ to define the incremental relative importance of distortion and entropy, the optimal entropy-constrained scalar quantizer (ECSQ) is the one which minimizes the objective function $J_X = D_X + \lambda H_X$ [12].

B. Quantization of exponential random variable

In this section, we review the features of an exponential random variable quantizer. We review the fact that the optimal ECSQ for exponential distribution is uniform, and define its stepsize and offset value in terms of the Lagrangian multiplier λ, and of the mean value of the random variable. These results are useful when studying MP atom quantization because the dead-zone subtracted modulus distribution of the matching pursuit atoms closely matches an exponential distribution. Formally, if the MP expansion selects atoms up to a threshold Θ, the Θ-subtracted modulus random variable M_Θ has an exponentially decreasing pdf [8]. Defining μ to be the mean value of M_Θ, and m to be the atom modulus, the pdf is

$$f_{M_\Theta}(x) = \mu^{-1} e^{-x/\mu}, \quad x = m - \Theta \geq 0, \mu > 0 \tag{8}$$

The key to analysis of the exponentially distributed random variable M_Θ lies in the fact that, under the condition that M_Θ has a value exceeding some fixed non-negative threshold t, the pdf of $(M_\Theta - t)$ is the same as the pdf of the original random variable M_Θ [12]. With respect to quantization, an immediate consequence of this property is that the optimal quantizer is uniform, i.e. all its steps have the same size Ω and reconstruction offset δ. The conditions for ECSQ optimality, together with the squared-error distortion and output entropy of the quantizer, are defined in [12]. For a given Lagrangian multiplier λ, the optimal stepsize Ω and offset δ of the quantizer are defined in terms of the mean value μ of the exponential distribution by the two following equations [12]

$$\lambda = -\frac{\partial D_{M_\Theta}(\Omega)}{\partial H_{M_\Theta}} = \frac{\ln(2)\ \left[(\Omega - \delta)^2 - \delta^2\right]}{\mu^{-1}\Omega} \tag{9}$$

and

$$\delta = \mu \left(1 - \frac{\mu^{-1}\Omega\ e^{-\Omega/\mu}}{1 - e^{-\Omega/\mu}}\right). \tag{10}$$

For small $\mu^{-1}\Omega$, from Taylor series expansion, we get $\delta \approx \Omega/2$ so that the optimal quantizer becomes a mid-quantizer, i.e. a quantizer for which the reconstructed value is the middle value of the quantization interval, and (9) becomes

$$\lambda = -\frac{\partial D_{M_\Theta}(\Omega)}{\partial H_{M_\Theta}} = \frac{\ln(2).\Omega^2}{6} \tag{11}$$

which conforms to the result derived from high-resolution analysis [4], valid for any pdf. The results reviewed in this section will be used to derive the optimal uniform quantizer in Sections III-C.2 and IV-A.

C. In-loop quantization of MP atom modulus

In this section, we consider the quantization of the atom moduli along the MP expansion process. We demonstrate that, due to the re-injection of the quantization error into the MP expansion loop, the atoms that are selected during the earlier MP iterations can be quantized more coarsely than the latter ones. This is because their quantization error is partly corrected by subsequent MP iterations. The benefit of the quantization error re-injection is taken into account to determine the uniform quantizer stepsize. The re-injection also motivates the design of a novel non-uniform quantization scheme. Non-uniformity is justified by the observation that atoms are roughly selected in decreasing order of magnitude and that, consequently, the earlier atoms are the largest ones. The benefit of re-injection is thus stronger for larger atom moduli, which suggests increasing the quantization stepsize with the atom magnitude.

C.1 Quantization error re-injection

Here, we consider the impact of quantization on the reconstructed MP expansion. As we consider independent quantization of the MP coefficients, the quantization distortion on the expansion is equal to the sum of the distortion on each coefficient. This distortion is different for *a posteriori* and *in-loop* quantization.

For *a posteriori* quantization, an entire set of atoms is extracted before quantization. Such scheme might be useful to adapt a single expansion to different transmission conditions. A variable subset of pre-selected atoms is then chosen and properly quantized to provide the best possible representation under different rate constraints. In that case, as mentioned in Section III-B, assuming an exponential distribution and independent quantization of the modulus, all quantized samples have the same distortion and entropy [12]. The optimal quantizer is uniform, and (9) gives its optimal quantizer stepsize for a given multiplier λ. Note that [2] proposes another quantizer for *a posteriori* quantization. Their quantizer is designed to encode atoms ordered in decreasing order of magnitude. Our framework is different as atoms are ordered to achieve high coding efficiency on the indices.

To improve coding efficiency, it is worthwhile to use the quantized atom modulus in the residue calculation [3]. Such *in-loop* quantization re-injects the quantization error in the MP expansion loop so that it can be corrected by subsequent iterations. *In-loop* quantization is particularly justified for conventional hybrid video coding whereby MP is used to expand a motion compensation prediction error [7]. In this framework, the signal to be expanded depends on a previously encoded reference frame. If the reference frame changes, the signal to be expanded in subsequent frames also changes. Hence, the synchronization between encoder and decoder prevents the exploitation of the flexibility offered by multiple quantizations of a single expansion when the reconstructed expansion is involved in a temporal prediction stage. In practice, multiple *a posteriori* quantizations of a single expansion are thus only useful when the expansion is not used for prediction purposes.

In the following, we analyze the evolution of the energy of the MP residue along the quantization expansion process in order to understand the impact of the re-injection on the distortion of the quantized moduli. Using the notations introduced in Section II, f is a given source vector in R^M and $\mathcal{R}_i f$ is the expansion residue after i MP iterations. At the initial step, $\mathcal{R}_0 f = f$ is the signal to be expanded. At step i, the algorithm selects the dictionary function φ_{k_i} that best matches $\mathcal{R}_i f$ and generates the residue for the next iteration. $\hat{\alpha}_i$ is defined to be the quantized value of $\alpha_i =< \mathcal{R}_i f, \varphi_{k_i} >$ and the residue based on the quantized modulus is $\mathcal{R}_{i+1} f = \mathcal{R}_i f - \hat{\alpha}_i \varphi_{k_i}$. Defining σ_i^2 as the i^{th} coefficient quantization distortion, and observing that $\mathcal{R}_i f - \alpha_i \varphi_{k_i}$ is orthogonal to φ_{k_i}, with $||.||$ referring to the quadratic norm in R^M, we have

$$||\mathcal{R}_{i+1} f||^2 = ||\mathcal{R}_i f||^2 - \alpha_i^2 + \sigma_i^2 \tag{12}$$

Without loss of generality, we can assume that at step i a fraction γ_i of the residual signal energy is captured by the MP expansion, i.e. $\alpha_i^2 = \gamma_i ||\mathcal{R}_i f||^2$. Then, we have

$$\begin{aligned}
||\mathcal{R}_1 f||^2 &= ||\mathcal{R}_0 f||^2 - \alpha_0^2 + \sigma_0^2 \quad with \;\; \alpha_0^2 = \gamma_0 ||\mathcal{R}_0 f||^2 \\
&= (1-\gamma_0)||\mathcal{R}_0 f||^2 + \sigma_0^2 \\
||\mathcal{R}_2 f||^2 &= ||\mathcal{R}_1 f||^2 - \alpha_1^2 + \sigma_1^2 \quad with \;\; \alpha_1^2 = \gamma_1 ||\mathcal{R}_1 f||^2 \\
&= (1-\gamma_0)(1-\gamma_1)||\mathcal{R}_0 f||^2 + (1-\gamma_1)\sigma_0^2 + \sigma_1^2 \\
||\mathcal{R}_{n+1} f||^2 &= \Pi_{i=0}^{n}(1-\gamma_i)||\mathcal{R}_0 f||^2 + \sum_{i=0}^{n} \left(\Pi_{j=i+1}^{n}(1-\gamma_j)\right) \sigma_i^2 \qquad (13)
\end{aligned}$$

In the above equation, the factor $\left(\Pi_{j=i+1}^{n}(1-\gamma_j)\right)$ multiplying σ_i^2 is a direct consequence of re-injecting the quantization error into the MP expansion loop. The factor $\left(\Pi_{j=i+1}^{n}(1-\gamma_j)\right)$, which is smaller than one, is specific to *in-loop* quantization, and reduces the impact of quantization error on the final distortion $||\mathcal{R}_{n+1} f||^2$. In the following, we consider Equation (13) to design the uniform and non-uniform *in-loop* optimal quantizer.

C.2 Uniform quantization

For uniform quantization and exponentially distributed coefficients, the distortion does not depend on the coefficient quantizer bin. So σ_i^2 in Equation (13) is the same for all coefficients. It is independent of i, and corresponds to the squared-error distortion of the quantizer defined in (6). For an exponential random variable, this distortion is defined in [12] as a function of the quantizer stepsize, and is denoted here by $D_{M_\Theta}(\Omega)$ for the Θ-subtracted modulus. When a total of N_Θ atoms are selected up to the stopping threshold Θ, the distortion $D_{UQ}(N_\Theta, \Omega)$ resulting from the *in-loop* uniform quantization of these atoms is the second term of the sum in (13) and can be written as

$$D_{UQ}(N_\Theta, \Omega) = \sum_{i=0}^{N_\Theta - 1} \left(\Pi_{j=i+1}^{N_\Theta - 1} (1 - \gamma_j) \right) \; D_{M_\Theta}(\Omega) \tag{14}$$

The entropy $H_{UQ}(N_\Theta, \Omega)$ for N_Θ atoms is N_Θ times the entropy of the exponential random variable corresponding to the Θ-subtracted modulus. This entropy is defined in [12] as a function of the quantizer stepsize, and is denoted here by $H_{\,M_\Theta}(\Omega)$. So, we have

$$H_{UQ}(N_\Theta, \Omega) = N_\Theta \ . \ H_{M_\Theta}(\Omega) \tag{15}$$

For a given number of atoms N_Θ selected up to the stopping threshold Θ, and a Lagrange multiplier λ, the R/D optimal quantizer stepsize Ω minimizes $D_{UQ}(N_\Theta, \Omega) + \lambda \ . \ H_{UQ}(N_\Theta, \Omega)$, i.e. it solves

$$\lambda = - \frac{\partial D_{UQ}(\Omega)}{\partial H_{UQ}} \tag{16}$$

Assuming that $\{\gamma_j\}$, or equivalently the ability of MP to capture the energy of the signal to be expanded, is not significantly affected by small variations of the quantizer stepsize, (16) becomes

$$\begin{aligned} \lambda &= -\frac{\sum_{i=0}^{N_\Theta - 1} \left(\Pi_{j=i+1}^{N_\Theta - 1} (1 - \gamma_j) \right)}{N_\Theta} \ . \ \frac{\partial D_{M_\Theta}(\Omega)}{\partial H_{M_\Theta}} \\ \lambda &= -\beta(N_\Theta) \ . \ \frac{\partial D_{M_\Theta}(\Omega)}{\partial H_{M_\Theta}} \end{aligned} \tag{17}$$

In the above equation, the ratio $\partial D_{M_\Theta} / \partial H_{M_\Theta}$ represents the change in distortion for a change in rate when quantizing the random variable M_Θ. This ratio is a function of the quantization stepsize Ω and is defined by (9) for an exponentially distributed random variable and by (11) at high resolution, i.e. for small Ω. The multiplication by a factor $\beta(N_{\,\Theta})$ that is smaller than one is due to re-injection, i.e. to the fact that the atom quantization error is partly corrected by subsequent MP iterations. It reflects the fact that, for a constant number of atoms N_Θ, re-injection reduces the impact of a change of rate, or equivalently of Ω, on the final distortion. As a consequence, for a given Lagrangian multiplier λ and under the reasonable assumption that the ability of MP to capture the signal energy is not significantly affected by small variations of the quantizer stepsize, the optimal quantization step is larger for *in loop* than for *a posteriori* quantization, where $\beta(N_{\,\Theta}) = 1$. This is obvious when considering small Ω whereby, inserting (11) into (17) results in

$$\lambda = \beta(N_\Theta) \ . \ \frac{\ln(2).\Omega^2}{6} \tag{18}$$

This clearly shows that for a given λ, the optimal Ω increases as $\beta(N_\Theta)$ decreases.

The developments provided in Appendix A approximate $\beta(N_\Theta)$ in terms of the ratio between the initial and final MP residual energy, i.e.

$$\beta(N_\Theta) \approx \sqrt{\frac{||\mathcal{R}_{N_\Theta} f||^2}{||\mathcal{R}_0 f||^2}} \tag{19}$$

Since an optimal expansion selects the atoms up to a stopping threshold, the final residual energy $||\mathcal{R}_{N_\Theta} f||^2$ is also the energy after all atoms larger than Θ have been selected. In that sense, the β factor in Equation (19) is a function of Θ. Let E_Θ be the residue energy after all atoms larger than Θ have been selected, and E_∞ be the initial residue energy. Then, (19) can be rewritten as

$$\beta(N_\Theta) \approx \sqrt{\frac{E_\Theta}{E_\infty}} \triangleq \beta'(\Theta) \tag{20}$$

and (17) becomes

$$\lambda = -\beta'(\Theta)\ .\ \frac{\partial D_{M_\Theta}(\Omega)}{\partial H_{M_\Theta}} \tag{21}$$

In practice, *in-loop* quantization requires the knowledge of Ω before starting the MP expansion. As a consequence, $\beta\,'(\Theta)$ has to be estimated ahead of the MP expansion. However, the final residual energy is only known once the expansion has been performed. Hence, we face a chicken-and-egg problem. This problem is considered in Section IV-A. Before moving to that, we first explain how to design the optimal non-uniform quantizer.

C.3 Non-uniform quantization

In the previous paragraph we have shown that re-injection affects the uniform quantizer stepsize through factor $\beta\,'(\Theta)$ in (21). Now we consider the design of the entropy constrained quantizer in presence of re-injection but in absence of uniformity constraint, which means that the quantizer can now be defined in terms of any sequence of stepsizes $\{\Omega_i\}_{i\geq 0}$, as defined in Section III-A. Designing the R/D optimal non-uniform quantizer consists of fixing its bin boundaries, or equivalently the sequence of stepsizes $\{\Omega_i\}_{i\geq 0}$, so that all bins have the same incremental benefit in distortion for a given incremental cost in rate, the ratio between them being defined by the Lagrangian multiplier λ.

Two observations justify the design of a non-uniform quantizer. Our first observation has to do with the fact that the quantization error of atoms that are selected in earlier MP iterations have a higher chance of being reduced in later MP iterations. This is reflected in (13) by the fact that the quantization distortion σ_i^2 is multiplied by a factor that decreases as i increases. Our second observation is that, as MP atoms are selected in decreasing order of magnitude, the largest atoms are selected in the initial iterations of the MP expansion. Combining these two observations, we conclude that the larger a coefficient, the more likely it has been selected in the earlier MP iterations, and the more chance its quantization error is reduced in subsequent iterations. This suggests a coarser quantization for larger coefficients, and shows that a non-uniform quantizer could be potentially useful in MP coding.

We now consider the design of the non-uniform quantizer. The main assumption underlying the design is the fact that atoms are selected mostly in decreasing order of magnitude. We use a recursive approach. At each step, given the lower boundary of a quantizer bin, the upper boundary is determined. The lower boundary of the first bin is the stopping threshold Θ. Let us now consider the selection of the upper boundary of the i^{th} quantizer bin, assuming its lower boundary is known. Let n_i denote the iteration index for which the selected atom modulus α_{n_i} equals the lower boundary of the i^{th} quantizer bin. According to this definition, $\alpha_{n_1} = \Theta$, and $\alpha_{n_{i+1}}$ denotes the upper boundary of the i^{th} quantizer bin. As atoms are selected in decreasing order of magnitude, for any $\alpha_{n_{i+1}} > \alpha_{n_i}$, the iteration index n_{i+1} is smaller than n_i, and the set of atoms selected between the n_{i+1}^{th} and n_i^{th} iterations of the MP expansion belong to $[\alpha_{n_i}, \alpha_{n_{i+1}}[$. Without loss of generality our problem is to find $\alpha_{n_{i+1}}$, or equivalently n_{i+1}, so that the best Lagrangian R/D trade-off is achieved for the quantization of the $(n_i - n_{i+1})$ atoms belonging to $[\alpha_{n_i}, \alpha_{n_{i+1}}[$. Letting N_Θ denote the total number of atoms selected by the MP expansion up to the Θ threshold, the distortion $D_i(N_\Theta, n_i, n_{i+1})$ due to the quantization of the $(n_i - n_{i+1})$ atoms belonging to the i^{th} bin $[\alpha_{n_i}, \alpha_{n_{i+1}}[$ with a stepsize $\Omega_i = \alpha_{n_{i+1}} - \alpha_{n_i}$ is an immediate consequence of (13) and can be written as:

$$D_i(N_\Theta, n_i, n_{i+1}) = \sum_{k=n_{i+1}+1}^{n_i} \left(\Pi_{j=k+1}^{N_\Theta-1}(1-\gamma_j)\right) \;.\; D_{M_\Theta}(\Omega_i) \tag{22}$$

so that, the distortion per atom is

$$\frac{D_i(N_\Theta, n_i, n_{i+1})}{n_i - n_{i+1}} = \frac{\sum_{k=n_{i+1}+1}^{n_i}\left(\Pi_{j=k+1}^{N_\Theta-1}(1-\gamma_j)\right)}{n_i - n_{i+1}} \;.\; D_{M_\Theta}(\Omega_i) \tag{23}$$

Similar to (21), given the Lagrange multiplier λ, the optimal i^{th} bin quantizer stepsize Ω_i solves

$$\lambda = -\frac{\sum_{k=n_{i+1}+1}^{n_i}\left(\Pi_{j=k+1}^{N_\Theta-1}(1-\gamma_j)\right)}{n_i - n_{i+1}} \;.\; \frac{\partial D_{M_\Theta}(\Omega_i)}{\partial H_{M_\Theta}} \tag{24}$$

For the $i^{th} bin$, denote $\beta_i(N_\Theta, n_i, n_{i+1})$ to be

$$\beta_i(N_\Theta, n_i, n_{i+1}) = \frac{\sum_{k=n_{i+1}+1}^{n_i}\left(\Pi_{j=k+1}^{N_\Theta-1}(1-\gamma_j)\right)}{n_i - n_{i+1}} \tag{25}$$

and rewrite Equation (24) as

$$\lambda = -\beta_i(N_\Theta, n_i, n_{i+1}) \;.\; \frac{\partial D_{M_\Theta}(\Omega_i)}{\partial H_{M_\Theta}} \tag{26}$$

where Ω_i is the width of the i^{th} bin. Given the lower bound of the bin, the above equation defines the upper bound as a function of λ. The developments provided in Appendix B approximate $\beta_i(N_\Theta, n_i, n_{i+1})$ as

$$\beta_i(N_\Theta, n_i, n_{i+1}) \approx \frac{||\mathcal{R}_{N_\Theta} f||^2}{||\mathcal{R}_{n_i} f||^2} \cdot \sqrt{\frac{||\mathcal{R}_{n_i} f||^2}{||\mathcal{R}_{n_{i+1}} f||^2}} = \frac{||\mathcal{R}_{N_\Theta} f||^2}{\sqrt{||\mathcal{R}_{n_i} f||^2 \cdot ||\mathcal{R}_{n_{i+1}} f||^2}} \tag{27}$$

This factor depends on the residual signal energies after respectively n_i, n_{i+1} and N_Θ MP iterations, with $N_\Theta > n_i > n_{i+1}$.

To design the non-uniform quantizer, we use an alternative interpretation of these energy values. We note that each of them is the energy of the residual signal after all atoms larger than a particular threshold have been selected. $||\mathcal{R}_{N_\Theta} f||^2$ measures the energy of the final residue, i.e. once all atoms larger than the stopping threshold Θ have been selected. $||\mathcal{R}_{n_i} f||^2$ and $||\mathcal{R}_{n_{i+1}} f||^2$ are measured once the atoms larger than respectively the lower and higher boundary of the i^{th} bin have been selected. Formally, let θ_i be the lower boundary of the i^{th} bin, i.e.

$$\theta_i = \Theta + \sum_{l=1}^{i-1} \Omega_l \tag{28}$$

Define E_θ to be the energy of the residue after all atoms larger than θ have been selected. E_Θ denotes the final residual energy. E_{θ_i} denotes the energy after all atoms larger than the lower boundary of the i^{th} bin have been selected. So, (27) can be written

$$\beta_i(N_\Theta, n_i, n_{i+1}) \approx \frac{E_\Theta}{\sqrt{E_{\theta_i} \cdot E_{\theta_{i+1}}}} \triangleq \beta_i'(\Theta, \theta_i, \theta_{i+1}) \triangleq \beta_i'(\Theta) \tag{29}$$

For notation convenience, $\beta_i'(\Theta, \theta_i, \theta_{i+1})$ is simply referred to as $\beta_i'(\Theta)$ in the remainder of the paper as this notation captures the dependency on Θ and on the index of the bin. Hence, (26) can be rewritten as

$$\lambda = -\beta_i'(\Theta) \cdot \frac{\partial D_{M_\Theta}(\Omega_i)}{\partial H_{M_\Theta}} \tag{30}$$

Equations (29) and (30) summarize the conclusions drawn from this section. First, for a given λ, each term Ω_i of the optimal quantizer stepsize sequence $\{\Omega_i\}_{i>0}$ has to solve Equation (30) with an appropriate $\beta_i'(\Theta)$ estimation. As mentioned in Section II, the stopping threshold Θ has to be chosen according to (3) to achieve R/D optimality. Second, based on (29), $\beta_i'(\Theta)$ can be estimated as a function of the final residual energy and of the energies of the residual signals obtained after all atoms larger than the lower and upper boundaries of the i^{th} bin have been selected. In practice, these energy values are only known once the expansion has been performed, which means that $\{\beta_i'(\Theta)\}_{i>0}$ can only be estimated a posteriori. However, as *in-loop* quantization performs the quantization along the expansion process, the stepsize sequence $\{\Omega_i\}_{i>0}$ has to be known ahead of the expansion. Hence, we face a chicken-and-egg problem. This problem together with the construction of the R/D optimal *in-loop* quantized expansion is considered in the next section.

IV. JOINTLY OPTIMAL ATOM SELECTION AND QUANTIZATION

For a given relative importance of rate and distortion, i.e. a given Lagrange multiplier λ, Sections II and III respectively show when to stop the iterative MP expansion, and how to quantize the atoms. This section combines these results to derive an explicit relation between the expansion stopping threshold, and the optimal stepsize or sequence of stepsizes of both the uniform and non-uniform quantizer. To simplify this relation, in Section IV-A we show that we can reasonably assume that the expected coding cost for the last selected atom is constant, i.e. does not depend on the Lagrangian cost λ, or equivalently on the reconstruction quality provided by the expansion. We also make the high-resolution assumption, which means that we replace the expressions derived for exponentially decreasing pdf by their formulations for small quantizer stepsize. In Section V-A, our simulations demonstrate that the high-resolution assumption does hold in practice. Then, we describe how to use the simplified relation in a video coding context. We observe empirically that the logarithm of the residual energy is a linear function of the logarithm of the stopping threshold. For non-uniform quantization, this observation allows for defining each quantizer stepsize proportional to the stopping threshold. The sequence of proportionality factors is derived from a set of recursive equations that are signal-independent, which means that the factors are constant and can be pre-computed. So, ultimately, the non-uniform quantizer design only consists of a few multiplications, which does not significantly increase complexity.

A. Uniform quantization: derivation, simplification, and utilisation of an R/D optimal relation between Θ and Ω

We consider uniform quantization and derive an explicit relation between the stopping threshold Θ and the uniform quantizer stepsize Ω for an R/D optimal MP expansion. We also simplify this relation and explain how to use it, both for isolated signals and for the successive frames of an hybrid MP video coder.

A.1 Derivation of the R/D optimal relationship between Θ and Ω

For uniform quantization, combining (3) and (21) removes the dependency on λ and defines the relation between the dead-zone threshold Θ, i.e. the threshold modulus below which it is worthless to select an atom, and the optimal quantizer stepsize Ω. Inserting (9) in this relation, the optimal stepsize Ω solves

$$\frac{\Theta^2 - \xi^2_{Q(.)}(\Theta)}{\Delta R_{index} + R_{Q(.)}(\Theta)} = \beta'(\Theta)\ \frac{\ln(2)\ \left[(\Omega - \delta(\Omega))^2 - \delta^2(\Omega)\right]}{\mu^{-1}\Omega} \tag{31}$$

where $\xi^2_{Q(.)}(\Theta)$ is the quantization error of an atom whose modulus is Θ, i.e. $\xi^2_{Q(.)}(\Theta) = \delta^2(\Omega)$ with $\delta(\Omega)$ defined by (10). The expected rate for the quantized modulus $R_{Q(.)}(\Theta) = \log_2 p_0$, with p_0 defined by (5). The multiplicative factor $\beta'(\Theta)$ is equal to one for *a posteriori* quantization but has to be estimated based on (20) for *in-loop* quantization.

A.2 Simplification of the Ω vs Θ relationship

Equation (31) cannot be easily used to determine the quantization stepsize Ω, since a number of parameters in (31) depend on the atom distribution statistics, which for *in-loop* quantization depend on the quantizer stepsize. Hence, we face a chicken-and-egg problem. In the following we first propose an iterative approach to solve this problem, and analyze the solution to validate a number of approximations that simplify (31). Then, we explain how to use the simplified relation in a video coding context, in a computationally efficient way.

In Equation (31), an appropriate choice of the threshold Θ is a rate-control issue. This is a problem similar to the selection of the quantization parameter QP for orthogonal transforms. Θ determines the quality, and indirectly the number of bits, of the expansion. The lower the threshold, the higher the signal representation quality and the bit-budget. For a given Θ, (31) provides the coder with the optimal quantizer stepsize. However, the equation depends on a number of parameters that have to be estimated, such as ΔR_{index}, μ and β'.

ΔR_{index} corresponds to the incremental cost in bits to define the sign and index of the additional atom. When predictive methods are used to code the indices of the atoms, ΔR_{index} depends on the atom distribution over the signal. As an example, for video coding applications, the atom positions are differentially encoded [1]. The parameter μ characterizes the shape of the exponential distribution of atom modulus and thus depends on the selected atoms. For *a posteriori* quantization this dependency is not a problem because the atoms are selected before being quantized. On the contrary *in-loop* quantization uses the quantized value in the residue calculation. As a consequence, it needs the quantizer stepsize Ω prior to the MP expansion, i.e. before μ and ΔR_{index} can be estimated from the selected atoms distribution. Moreover, for *in-loop* quantization, $\beta\,'(\Theta)$ is not equal to one and is estimated from the energy captured by the atoms. Thus, it also depends on the selected atoms.

To solve this chicken-and-egg problem, we propose an iterative approach as follows. Initial μ and $\Delta R_{\,index}$ estimations are obtained by selecting unquantized atoms until the stopping criterion is met. $\Delta R_{\,index}$ is estimated by the average index and sign coding cost of the selected atoms. Given this initial expansion, $\beta\,'(\Theta)$ is estimated based on (20) and Ω is computed from (31). Quantized atoms are then selected and better estimations of μ, ΔR_{index} and $\beta'(\Theta)$ are computed. Convergence is most often achieved in a single iteration. However, this iterative approach is computationally expensive and impractical. Its purpose is to provide a reference solution to validate our proposed simplification of (31) to be described shortly.

Table III presents results derived from the iterative approach. Four different motion residual frames have been expanded using three different Θ thresholds. The “Foreman”, “Silent”, and “Mobile” sequences are considered, either in QCIF or CIF format. The uniform quantizer stepsize Ω has been derived, together with an estimate of the parameters in (31), using the

Seq. Fr.	Θ	μ	β'^{-1}	Ω	$\mu^{-1}\Omega$	δ	R_{last}	$N_{at.}$	dB
QFor. 60	60	18	1.1	43	2.4	13	18.6	50	32
QFor. 60	30	15	1.3	22	1.5	9	18.0	396	35
QFor. 60	10	13	2.6	10	0.8	4	18.0	1689	42
QSil. 60	60	12	1.1	45	3.7	12	19.0	16	32
QSil. 60	30	10	1.3	23	2.3	8	18.5	258	34
QSil. 60	10	10	2.4	11	1.1	4	17.7	1925	41
For. 40	60	18	1.0	42	2.4	13	18.4	13	34
For. 40	30	8	1.1	22	2.7	7	18.4	427	36
For. 40	10	8	2.0	10	1.2	4	17.9	4987	42
Mob. 40	60	25	1.3	48	1.9	17	17.8	850	29
Mob. 40	30	20	2.0	28	1.4	11	17.8	3776	33
Mob. 40	15	18	3.2	18	1.0	8	17.0	9180	37

TABLE III

Motion residual frames have been expanded using three different Θ thresholds. The atom statistical distribution parameter μ and the uniform quantizer stepsize Ω have been derived from (31), using the iterative approach described in the text.

iterative approach described above. An immediate observation is that the estimated cost in bits of the last selected atom, i.e. $R_{last} = \Delta R_{index} + R_{Q(.)}(\Theta)$, does not significantly depend on the stopping threshold Θ, or on the video sequence. This observation, valid for motion residual image representation, can likely be extended to MP signal coding in general. Indeed, to provide a sparse representation, i.e. a representation with few significant coefficients, MP decomposes a signal over a highly redundant dictionary and devotes a large amount of bits to the atom indices. Thus, $R_{Q(.)}(\Theta)$, which depends directly on the modulus quantization, is not the dominant part of R_{last}. Moreover, as shown in Table III, the finer the quantization, the higher the representation quality and the number of atoms. A larger number of atoms is expected to improve the coding efficiency for indices differential encoding methods, and compensate for the finer and more expensive quantization.

These observations can be used to simplify (31). Assuming R_{last} to be a constant parameter and considering small $\mu^{-1}\Omega$ values, i.e. making the high-resolution assumption, (31) becomes

$$\frac{\Theta^2 - \delta^2}{R_{last}} = \beta'(\Theta)\ .\ \ln(2).\Omega^2/6 \tag{32}$$

From small $\mu^{-1}\Omega$, as stated in Section III-B, $\delta = \Omega/2$ and hence (32) becomes

$$\frac{\Omega}{\Theta} = \sqrt{\frac{1}{\beta'(\Theta)\ln(2).R_{last}/6 + 1/4}} \tag{33}$$

According to our simulations, for hybrid video coding $R_{last} \in [17 \ldots 19]$, and we have

$$\frac{\Omega}{\Theta} \approx 0.66\ .\ \sqrt{\beta'(\Theta)^{-1}} \tag{34}$$

For $\beta'(\Theta) = 1$, this result is close to the 0.6 empirical ratio found in [8]. The analytical derivation clarifies assumptions under which the result is valid, and refines it when the impact of re-injection becomes significant, i.e. when $\beta'(\Theta) << 1$.

Note that with $\beta'(\Theta) = 1$, Equation (33) also defines the optimal uniform quantizer for *a posteriori* quantization when atom moduli are coded independently, e.g. when efficient entropy coding of the atoms positions defines the atoms coding order. As mentioned in Section III-C.1, *a posteriori* quantization is useful when variable subsets of a single set of pre-selected atoms are chosen to provide the best possible representations under different rate constraints, i.e. for different stopping thresholds. In practice, for a given stopping threshold Θ, all the atoms with a modulus larger than Θ are selected within the pre-computed MP expansion, and are quantized based on (33).

Before explaining how to use (34) for *in-loop* quantization, it is interesting to notice that the high-resolution assumption makes the quantizer design independent of the exponential decreasing characteristic of the atom modulus distribution. The high resolution assumption can be verified a posteriori by comparing the value of Ω estimated by equation (34) with the value of Ω presented in Table III. The latter solves Equation (31) and results from an estimation of the μ and R_{last} parameters. In Section V-A, the comparison between quantizers designed with and without the high-resolution assumption definitely demonstrates that the assumption does hold in practice.

A.3 Iterative $\beta'(\Theta)$ estimation for isolated signals

For a given Θ, Equation (34) specifies the proper quantization stepsize Ω, provided $\beta'(\Theta)$ is known. In practice computing $\beta'(\Theta)$ is problematic. In particular, recall from (20) that $\beta'(\Theta)$ depends on E_Θ, the residual energy after all atoms larger than Θ have been selected. E_Θ can only be computed after the $\beta'(\Theta)$ dependent expansion has been completed, resulting in a chicken-and-egg problem. Even though $\beta'(\Theta)$ is needed for the expansion itself, one approximate solution is to assume $\beta'(\Theta) = 1$ in computing an initial expansion, and use the resulting E_Θ to refine $\beta'(\Theta)$ for computing another expansion. However, this solution is computationally expensive. In practice, the initial expansion can be computed with an approximate dictionary, as suggested in [9], [10], but it still remains more complex than a single expansion.

A.4 $\beta'(\Theta)$ estimation based on a previously encoded frame in MP video coding

We consider uniform quantization for a specific application, which is the MP expansion of the motion compensation prediction error obtained for hybrid video coding. The hybrid MP video codec we study is the one proposed in [7], [8]. The frame to encode is first predicted by motion compensation of a previously encoded frame. MP expansion is then used to expand and encode the prediction error frame. In this context, a computationally attractive alternative is to estimate E_Θ from the previous frame. As the parameter Θ is likely to change between successive frames, we have to analyze the dependency of E_Θ on Θ. Fig. 1 depicts the final residual energy of four different video frames as a function of Θ. A logarithmic scale is used for both axes. For small Θ, corresponding to higher bit rates, the relation is linear. As Θ increases, E_Θ converges to the original signal

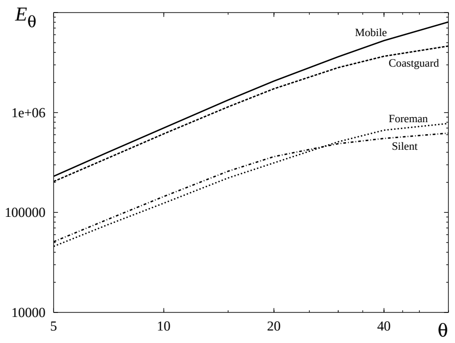


Fig. 1. Residual energy E_Θ as a function of the stopping threshold Θ. Logarithmic scales are used for both axes. Each curve corresponds to the 60^{th} frame of a video sequence. Silent and Foreman have been encoded in QCIF format. Mobile and Coastguard are CIF sequences.

energy. In this case, few atoms are selected and re-injection can be neglected, i.e. $\beta'(\Theta) = 1$. The interesting part of the graph is the linear part. Denote ρ to be its angular coefficient. As axes are in logarithmic scales, given two values θ and θ' that lie in the linear part of the graph, we have

$$\frac{E_{\theta'}}{E_\theta} = \left(\frac{\theta'}{\theta}\right)^\rho \tag{35}$$

From Fig. 1, ρ can be reasonably assumed to be sequence independent. In fact, it is close to 1.5 for all sequences. This observation provides a convenient method to estimate the residual energy, and consequently the β' factor, as Θ changes between successive frames. Specifically, denote Θ^j and $E^j_{\Theta^j}$ to be the stopping threshold and the final residual energy for the j^{th} frame respectively. Let Θ^{j+1} be the Θ-parameter chosen for the $(j+1)^{th}$ frame, and E^{j+1}_∞ be the original energy of the $(j+1)^{th}$ displaced frame difference. Then, $\beta'(\Theta^{j+1})$ can be estimated based on (20) and (35) with $\rho = 1.5$,

$$\beta'(\Theta^{j+1}) \approx \sqrt{\frac{E^j_{\Theta^j}\ .\ (\Theta^{j+1}/\Theta^j)^\rho}{E^{j+1}_\infty}} \tag{36}$$

In conclusion, the R/D optimal expansions of the successive displaced frame differences of a video codec are computed as follows. Let Θ^j and $\beta'(\Theta^j)$ denote the Θ parameter and β' factor for the j^{th} frame. For the first frame, $\beta'(\Theta^1)$ is set to one. Hence, the optimal quantizer stepsize for the j^{th} frame is computed based on (34), and the j^{th} displaced frame is expanded. The final residual energy $E^j_{\Theta^j}$ is then used to estimate the β'-factor for the next frame based on (36).

B. Non-uniform quantization: β'-sequence estimation

Fundamentally, our analytical derivation in Section III-C.3 has shown that, because of the quantization error re-injection, the uniform quantization is not optimal. In this section, we explain how to compute the R/D optimal MP expansion with *in-loop* non-uniform quantization. We combine (3) and (30) to define the relation between Θ and the sequence of stepsizes. Each element Ω_i of the optimal stepsize sequence $\{\Omega_i\}_{i>0}$ solves

$$\frac{\Theta^2 - \xi^2_{Q(.)}(\Theta)}{\Delta R_{index} + R_{Q(.)}(\Theta)} = \beta'_i(\Theta)\ \frac{\ln(2)\ \left[(\Omega_i - \delta(\Omega_i))^2 - \delta^2(\Omega_i)\right]}{\mu^{-1}\Omega_i} \tag{37}$$

Similar to (31), the above equation can be simplified to result in

$$\frac{\Omega_i}{\Theta} \approx 0.66\ .\ \sqrt{\beta'^{-1}_i(\Theta)}, \quad with\ \ \beta'_i(\Theta) \approx \frac{E_\Theta}{\sqrt{E_{\theta_i}\ .\ E_{\theta_{i+1}}}} \tag{38}$$

For a given Θ, this set of equations specifies the proper sequence of quantization stepsizes $\{\Omega_i\}_{i>0}$, provided $\{\beta'_i(\Theta)\}_{i>0}$ is known. However, the energy values needed to estimate $\{\beta'_i(\Theta)\}_{i>0}$ are only known after the expansion has been completed. In this section, we first propose two methods to address this chicken-and-egg problem for any given signal. We then comment and discuss their practical use in a video coding context. The section concludes on the observation that a constant sequence of factors provides a good approximation of $\{\beta'_i(\Theta)\}_{i>0}$. It explains how to compute the constant sequence of factors, and observe that the use of a constant sequence of factors virtually removes the quantizer design cost.

B.1 β'-sequence estimation for isolated signals

The first approach to estimate $\{\beta'_i(\Theta)\}_{i>0}$ is derived directly from (38) and proceeds iteratively. In the following it is referred to as the Iterative Sequence Estimation (ISE) method. The ISE method can be formalized as follows. For a given Θ, let $\{\beta'^{j-1}_i(\Theta)\}_{i>0}$ be the β'-sequence derived for the $(j-1)^{th}$ iteration with $j > 0$. The process is initialized with $\beta'^0_i(\Theta) = 1$ for all i. At the j^{th} iteration, $\{\Omega^j_i\}_{i>0}$ is computed from (38), using $\{\beta'^{j-1}_i(\Theta)\}_{i>0}$, and the j^{th} signal expansion is performed with *in-loop* quantization. In order to process the next iteration, the sequence $\{\beta'^j_i(\Theta)\}_{i>0}$ is then estimated based on (38), in terms of the final residual energy and of the residual energy after all the atoms larger than the boundaries of the i^{th} quantization bin have been selected. These energy values are based on the j^{th} expansion, and are denoted by E^j_Θ, $E^j_{\theta_i}$, and $E^j_{\theta_{i+1}}$ respectively. They are computed by accumulating, on a quantizer bin basis, the atom contribution to the residue energy decrease. The contribution of the k^{th} atom is $\alpha^2_k - (\alpha_k - \hat{\alpha}_k)^2$, where α_k and $\hat{\alpha}_k$ are respectively the non-quantized and quantized atom modulus. Let ΔE^j_i denote the sum of the contributions of atoms of the j^{th} expansion whose modulus lie in the i^{th} quantizer bin, i.e.

$$\Delta E^j_i = \sum_{\alpha_x \in [\theta^j_i, \theta^j_{i+1}]} \left(\alpha^2_x - (\alpha_x - \hat{\alpha}_x)^2\right), \qquad i > 0\ \ and\ \ \theta^j_i = \Theta + \sum_{l=1}^{i-1} \Omega^j_l \tag{39}$$

By definition of $E^j_{\theta_i}$, we have

$$E^j_{\theta_i} = E^j_\Theta + \sum_{l=1}^{i-1} \Delta E^j_l \tag{40}$$

and, based on the second part of (38), we have

$$\beta'^{j}_{i}(\Theta) \approx \frac{E^j_\Theta}{\sqrt{(E^j_\Theta + \sum_{l=1}^{i-1} \Delta E^j_l) \ . \ (E^j_\Theta + \sum_{l=1}^{i} \Delta E^j_l)}} \tag{41}$$

In practice, we have observed the ISE process to converge after a single iteration. We note from (41) that $\beta'^{j}_{i}(\Theta)$ is smaller than one and decreases as the bin index i increases. Based on (38), we conclude that Ω_i increases with the bin index i.

To second approach, called Recursive Sequence Computation (RSC), derives the β'-sequence based on (35), which states that for any θ_k and θ_l chosen in the set $\{\Theta, \{\theta_i\}_{i>0}\}$, provided the residual energies E_{θ_l} and E_{θ_k} are small compared to the initial energy E_∞,

$$E_{\theta_k} \approx E_{\theta_l} \ . \ (\theta_k/\theta_l)^\rho \tag{42}$$

Inserting (42) in (38) results in

$$\beta'_i(\Theta) \approx \sqrt{\frac{E_\Theta}{E_{\theta_i}} \cdot \frac{E_\Theta}{E_{\theta_{i+1}}}} \approx \sqrt{\left(\frac{\Theta}{\theta_i}\right)^\rho \cdot \left(\frac{\Theta}{\theta_{i+1}}\right)^\rho} \tag{43}$$

and inserting (43) in (38), we have

$$\frac{\Omega_i}{\Theta} \approx 0.66 \ . \ \left(\frac{\theta_i}{\Theta}\right)^{\rho/4} . \left(\frac{\theta_{i+1}}{\Theta}\right)^{\rho/4} \ , \tag{44}$$

provided the bin boundaries θ_i and θ_{i+1} are such that E_{θ_i} and $E_{\theta_{i+1}}$ are small in comparison with the initial energy E_∞. In the above equation, by definition, $\theta_i = \Theta + \sum_{l=1}^{i-1} \Omega_l$, which only depends on $\{\Omega_l\}_{l<i}$. Hence, (44) can be written as

$$\frac{\Omega_i}{\Theta} \approx 0.66 \ . \ \left(1 + \sum_{l=1}^{i-1} \frac{\Omega_l}{\Theta}\right)^{\rho/4} . \left(1 + \sum_{l=1}^{i-1} \frac{\Omega_l}{\Theta} + \frac{\Omega_i}{\Theta}\right)^{\rho/4} \ , \quad i > 0. \tag{45}$$

This non-linear set of equations can be solved recursively and independently of Θ to compute the sequence $\{\Omega_i/\Theta\}_{i>0}$. We define the χ-sequence $\{\chi_i\}_{i>0}$ by

$$\chi_i \triangleq (0.66)^{-1} \ . \ \Omega_i/\Theta \ , \quad i > 0, \tag{46}$$

The χ-sequence obtained from numerical solution to (45) with $\rho = 1.5$ is shown in Table V(a). Note that χ_i is entirely independent of the particular video or frame being processed, and only relies on the validity of the assumption in (42). Combining (46), (43) and (44) we get $\chi_i \approx \sqrt{\beta'^{-1}_i(\Theta)}$. As χ_i is computed independently of Θ, this would seem to imply that $\beta'_i(\Theta)$ does not depend on Θ either. However, (42), and consequently (43), are only valid if the residual energies E_{θ_i} and $E_{\theta_{i+1}}$ are small in comparison with the initial energy E_∞. As a consequence, the recursive sequence computation provides χ_i values that are good approximations of $\sqrt{\beta'^{-1}_i(\Theta)}$ only if E_{θ_i} and $E_{\theta_{i+1}}$ are small in comparison with the initial energy

E_∞. Using the second part of (38), this condition can be expressed in terms of $\beta_i'(\Theta)$, and we can conclude that χ_i provides a good approximation of $\sqrt{\beta'^{-1}_i(\Theta)}$ only if

$$\beta_i'(\Theta) \approx \sqrt{\frac{E_\Theta}{E_{\theta_i}} \cdot \frac{E_\Theta}{E_{\theta_{i+1}}}} > \frac{E_\Theta}{E_\infty} \tag{47}$$

Hence, χ_i provides a good approximation of $\sqrt{\beta'^{-1}_i(\Theta)}$ only if

$$\chi_i \approx \sqrt{\beta'^{-1}_i(\Theta)} < \sqrt{\frac{E_\infty}{E_\Theta}} \tag{48}$$

When this condition is not valid, i.e. when $E_{\theta_i} \approx E_{\theta_{i+1}} \approx E_\infty$, (38) results in

$$\beta_i'(\Theta) \approx \sqrt{\frac{E_\Theta}{E_{\theta_i}} \cdot \frac{E_\Theta}{E_{\theta_{i+1}}}} \approx \frac{E_\Theta}{E_\infty} \tag{49}$$

In conclusion, $\sqrt{\beta'^{-1}_i(\Theta)}$ can be approximated either by χ_i when $\chi_i < \sqrt{E_\infty/E_\Theta}$, or by $\sqrt{E_\infty/E_\Theta}$ when $\chi_i > \sqrt{E_\infty/E_\Theta}$. So, $\{\sqrt{\beta'^{-1}_i(\Theta)}\}_{i>0}$ can be approximated by a truncated version of the χ-sequence. Let l be the largest index such that $\chi_l < \sqrt{\frac{E_\infty}{E_\Theta}}$, and define the truncated sequence $\{\tau_i(\Theta)\}_{i>0}$ to approximate $\{\sqrt{\beta'^{-1}_i(\Theta)}\}_{i>0}$ as follows:

$$\begin{aligned}
\tau_i(\Theta) &= \chi_i, \quad 0 < i < l \\
\tau_i(\Theta) &= \sqrt{\frac{E_\infty}{E_\Theta}}, \quad i > l+2 \\
\tau_l(\Theta) &= \chi_l - 0.2 \\
\tau_{l+1}(\Theta) &= 2.\tau_l(\Theta)/3 + \tau_{l+3}(\Theta)/3 \\
\tau_{l+2}(\Theta) &= \tau_l(\Theta)/3 + 2.\tau_{l+3}(\Theta)/3
\end{aligned} \tag{50}$$

The τ-sequence is equal to the χ-sequence when (48) is met and tends to $\sqrt{E_\infty/E_\Theta}$ beyond that. Three samples are interpolated to insure a smooth transition between these two regions. Note that the truncation depends on E_Θ, which can only be computed after the expansion has been completed, resulting in a chicken-and-egg problem. One approximate solution is to assume $\tau_i(\Theta) = \chi_i$, $\forall i$ in computing an initial expansion, in order to determine E_Θ, which is then used to refine $\{\tau_i(\Theta)\}_{i>0}$ for a subsequent expansion.

Tables IV and V compare the estimates of $\sqrt{\beta'^{-1}_i(\Theta)}$ based on the two approaches described above. Table IV presents the estimation based on the ISE approach. In comparison, Table V refers to the RSC approach, which approximates $\{\sqrt{\beta'^{-1}_i(\Theta)}\}_{i>0}$ by the τ-sequence defined by (50). Table V(a) provides the Θ independent χ-sequence generated with $\rho = 1.5$, while Tables V(b) and (c) present the truncated τ-sequence corresponding to the $\sqrt{E_\infty/E_\Theta}$ values measured for the expansions analyzed in Table IV. We note the close match between the two approaches. Both are thus good candidates to derive the factors to be

Θ	dB	$\sqrt{\beta'^{-1}_1(\Theta)}$	$\sqrt{\beta'^{-1}_2(\Theta)}$	$\sqrt{\beta'^{-1}_3(\Theta)}$	$\sqrt{\beta'^{-1}_4(\Theta)}$	$\sqrt{\beta'^{-1}_5(\Theta)}$
40	34	1.0	1.1	1.1	1.1	1.1
20	37	1.2	1.5	1.6	1.6	1.6
10	41	1.2	1.7	2.1	2.4	2.5
5	45	1.2	1.8	2.5	3.1	3.7

(a) Foreman QCIF, 40^{th} frame (32 dB when $\Theta = \infty$).

Θ	dB	$\sqrt{\beta'^{-1}_1(\Theta)}$	$\sqrt{\beta'^{-1}_2(\Theta)}$	$\sqrt{\beta'^{-1}_3(\Theta)}$	$\sqrt{\beta'^{-1}_4(\Theta)}$	$\sqrt{\beta'^{-1}_5(\Theta)}$
40	31	1.1	1.4	1.5	1.6	1.6
20	35	1.2	1.7	2.1	2.3	2.4
10	40	1.3	1.9	2.6	3.2	3.7
5	45	1.3	2.0	3.0	4.0	5.0

(a) Mobile CIF, 40^{th} frame (27 dB when $\Theta = \infty$).

TABLE IV

β'-SEQUENCE PARAMETERS ESTIMATED USING THE ISE METHOD FOR DIFFERENT Θ THRESHOLDS.

χ_1	χ_2	χ_3	χ_4	χ_5	χ_6	χ_7	χ_8
1.2	1.9	2.8	3.9	5.2	6.4	7.7	8.8

(a) Reference χ-sequence, computed with $\rho = 1.5$.

Θ	$\sqrt{\frac{E_\infty}{E_\Theta}}$	$\tau_1(\Theta)$	$\tau_2(\Theta)$	$\tau_3(\Theta)$	$\tau_4(\Theta)$	$\tau_5(\Theta)$
40	1.2	1.0	1.1	1.1	1.2	1.2
20	1.7	1.1	1.3	1.5	1.7	1.7
10	2.8	1.2	1.7	2.1	2.4	2.8
5	4.4	1.2	1.9	2.8	3.7	3.9

(b) Foreman QCIF.

Θ	$\sqrt{\frac{E_\infty}{E_\Theta}}$	$\tau_1(\Theta)$	$\tau_2(\Theta)$	$\tau_3(\Theta)$	$\tau_4(\Theta)$	$\tau_5(\Theta)$
40	1.5	1.0	1.2	1.4	1.5	1.5
20	2.5	1.2	1.7	2.1	2.3	2.5
10	4.5	1.2	1.9	2.8	3.7	3.9
5	7.8	1.2	1.9	2.8	3.9	5.2

(c) Mobile CIF .

TABLE V

REFERENCE χ SEQUENCE AND τ-SEQUENCES COMPUTED TO APPROXIMATE THE β-SEQUENCE BASED ON THE RSC METHOD FOR DIFFERENT Θ THRESHOLDS.

used in (38). ISE is general in the sense that it does not rely on the validity of Equation (42), but in practice, requires two signal expansions. RSC, on the other hand, relies on a constant pre-computed χ-sequence. Similar to ISE, the truncation of this sequence requires two expansions of the same signal for RSC. However, in Section V-B our simulations demonstrate that, in practice, using the constant non-truncated χ-sequence provides a very good quality/complexity trade-off. This is because the τ- and χ-sequences only differ in high order bins, i.e. the ones that contain few atoms. Using (38) with $\sqrt{\beta'^{-1}_i(\Theta)} \approx \chi_i$ results in

$$\Omega_i = 0.66\ \chi_i\ \Theta \tag{51}$$

which indicates that Ω_i is merely a scaled version of Θ, with χ_i shown in Table V(a). Our experiments have validated the approximation made in equation (51) for different spatial resolutions (CIF and QCIF) and sequence contents (”Coastguard”,”Silent”,”Foreman”, and ”Mobile”). Moreover, the fact that the χ-sequence is signal independent means that this sequence can be pre-computed, which drastically reduces the non-uniform quantizer design computational complexity based on RSC. In the following, we refer to the RSC method using the non-truncated χ-sequence as reduced complexity RSC, or RC-RSC for short. In the next section, we discuss the specificities of both the ISE and RSC approaches in a video coding context.

B.2 β'-sequence estimation along a video sequence

When encoding entire video sequences, it is computationally attractive to use the expansion computed on the previous frame to estimate the β'-sequence for the current frame, thus avoiding multiple signal expansions. However, in practice Θ is chosen to target a given rate or quality, and is likely to change between successive frames. This results in a mismatch between the parameters computed based on the expansion of the $(j-1)^{th}$ frame and the one needed to estimate the β'-sequence for the j^{th} frame.

For this reason, the ISE method used in the video coding experiments presented in Section V-B is identical to the ISE method described in Section IV-B.1 for isolated signals, i.e. each motion prediction error signal is considered as an isolated signal. Specifically, for each motion prediction error and given the quality-dependent threshold Θ, an initial expansion is performed with a quantizer defined by $\Omega_i = 0.66.\Theta$. Given the initial expansion, the β'-sequence is then estimated based on (41), which defines the quantizer used for an additional expansion. This method is expensive from a computational point of view, but it provides a reference for comparison with the simplified RSC method described in the following.

The RSC method can be extended to video sequences in a much more efficient way than the ISE approach. For the RSC method, the previous frame expansion can be exploited to estimate the current $\beta\,'$-sequence. Letting Θ^j and $E^j_{\Theta^j}$ denote the stopping threshold and the final residual energy for the j^{th} frame respectively, the final residual energy of frame $(j+1)$ is estimated based on (35) by $E^j_{\Theta^j}(\Theta^{j+1}/\Theta^j)^\rho$. Recall that Θ^j and Θ^{j+1} are rate control related parameters, and hence are known prior to the expansion. Once the final residual energy of frame $(j+1)$ is estimated, (50) is used to estimate the τ-sequence, and the Ω-sequence for the $(j+1)^{th}$ expansion. This RSC approach is convenient to implement and takes the variation of Θ between frames into account. Specifically, RSC relies on the assumption that (42) holds, and requires the knowledge of a particular value of ρ to compute the χ-sequence. From Fig. 1, we observe that $\rho = 1.5$ is a reasonable assumption for video coding applications. Moreover, as mentioned in the previous section, our simulations have shown that, in practice, using the non-truncated χ-sequence rather than the τ-sequence is not only attractive from a computational complexity point of view, but also provides similar coding performance. Thus, for a video sequence, RC-RSC provides the most favorable trade-off between coding performance and complexity, as compared to RSC or ISE.

B.3 Conclusion and summary of non-uniform quantization results

In this section, we have proposed two practical schemes to design the non-uniform quantizer. We have shown that each stepsize Ω_i is proportional to the stopping threshold Θ, and that the proportionality factor can be derived either from an initial expansion of the current frame (ISE approach), or from the appropriate truncation of a constant sequence of factors (RSC method). The experiments presented in Section V-B indicate that these two approaches achieve identical performances. It proves that the optimal sequence of proportionality factors that is estimated by the ISE approach is well approximated by the RSC method. Moreover, these experiments also show that omitting the truncation (RC-RSC method), and thus using the constant χ-sequence rather than the τ-sequence, results in negligible performance penalty, so that in practice each quantizer stepsize can be defined as a scaled version of the stopping threshold Θ. From a computational complexity point of view, using the constant χ-sequence is attractive as it virtually removes the non-uniform quantizer design cost.

In conclusion, we recommend the use of the RC-RSC method, which defines the non-uniform quantizer based on Equation (51), i.e. for the studied MP video coding application, we have

$$\Omega_i = 0.66 \ \chi_i \ \Theta \tag{52}$$

Remembering that the 0.66 factor appearing in (52) has been derived based on (33), the general form of Equation (52) can be written

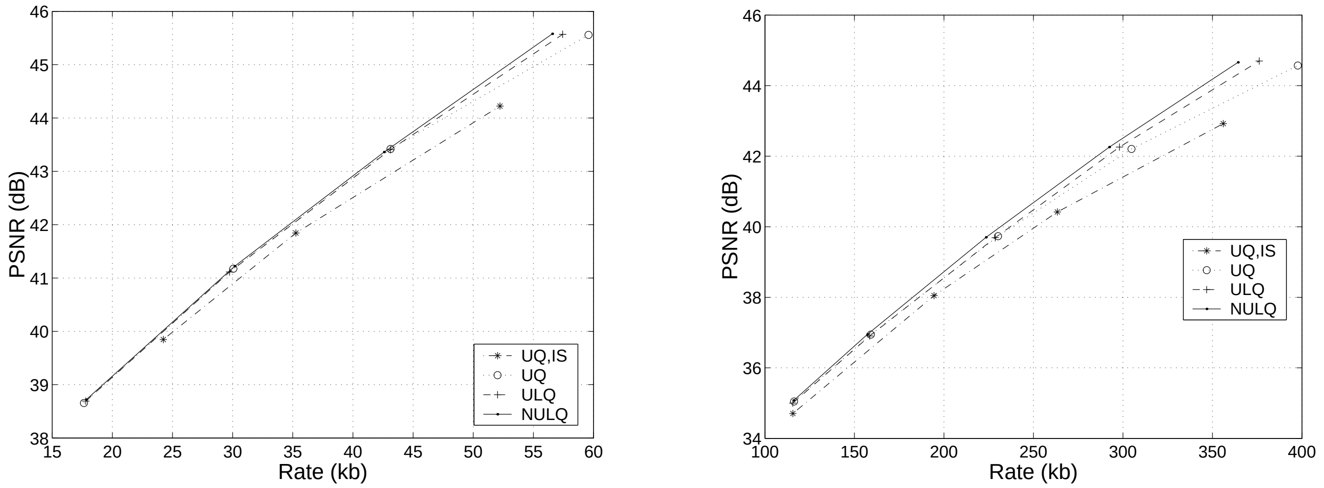
$$\Omega_i = \sqrt{\frac{6}{\ln(2).R_{last}}} \ \chi_i \ \Theta \tag{53}$$

where R_{last} denotes the expected rate for the last selected atom.

In (52) or (53)), $\{\chi_i\}_{i>0}$ solves the set of non-linear equations defined by (45), and written as

$$\chi_i \approx 0.66 \ . \ \left(1+\sum_{l=1}^{i-1}\chi_l\right)^{\rho/4} . \left(1+\sum_{l=1}^{i-1}\chi_l+\chi_i\right)^{\rho/4} \quad , \qquad i>0. \tag{54}$$

This non-linear set of equations can be solved recursively, and only depends on the ρ parameter defined by Equation (35). For video coding applications, we have shown in Figure 1 that $\rho = 1.5$ provides a reasonable and sequence-independent estimation of ρ, and Table V has provided the χ-sequence solving (54) with $\rho = 1.5$. In Appendix C, additional results demonstrate that the definition of ρ in Equation (35) is relevant for generic 1-D signals and dictionaries, which shows that the RC-RSC method can be extended to other MP coding applications than video coding.



(a) Foreman QCIF, fr.60.

(a) Mobile CIF, fr.40.

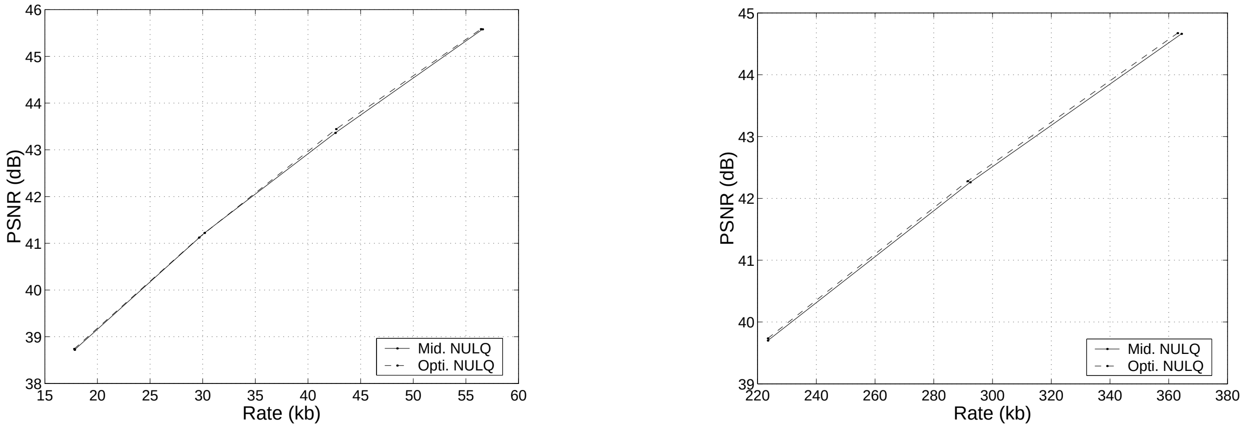
Fig. 2. DFD Rate/Distortion curves for a number of mid-quantizers. UQ,IS corresponds to [10]. It uses the uniform quantizer defined by Equation (34) with $\beta' = 1$. UQ uses the same uniform mid-quantizer, but the search goes on until no atom smaller than the threshold can be found in any part of the image. ULQ is the uniform mid-quantizer that takes the quantization error re-injection into account, i.e. $\beta \neq 1$. NULQ is the close-to-optimal non-uniform mid-quantizer.

V. RESULTS: APPLICATION TO VIDEO CODING

In this section, we consider the MP expansion of the motion compensation prediction error of the conventional hybrid MP video codec used in [7], [8]. We analyze the performance of the quantizers and atom selection criterion described above. As expected, we show that the non-uniform quantization of atoms selected up to a given threshold achieves the best rate-distortion trade-offs. The encoding of isolated Displaced Frame Difference (DFD) expansions is first considered. We demonstrate the validity of the approximations made to derive (38) from (37). We also motivate our work by comparing the coding efficiency of two MP video coders. The first one encodes the atoms in the order defined in [1], so that the differential coding of the atoms positions provides high entropy gains, and uses the uniform independent scalar modulus quantization method proposed in this paper to encode the atoms moduli. On the contrary, the second one, which is defined in Section V-A, encodes the atoms in decreasing order of magnitude, so that the atoms moduli can be differentially encoded. Our simulations show that the first approach, which is the one corresponding to the framework envisioned in this paper, achieves higher coding efficiency than the second one. Finally, the encoding of entire video sequences quantifies the benefit of both our uniform and non-uniform quantization methods, and confirms the relevance of deriving the β' parameter from the previous frame, and the β'-sequence from the constant χ-sequence.

A. Isolated signal case: DFD atom selection and quantization

Fig. 2 compares the rate/distortion performance of a number of uniform and non-uniform mid-quantizers for the 60 th and 40^{th} displaced frame difference of the Foreman and Mobile sequences. For all quantizers, the original displaced frame differences are identical. In all cases, for complexity reasons, the high dimensional motion residual image is split into overlapping blocks



(a) Foreman QCIF, fr.60.

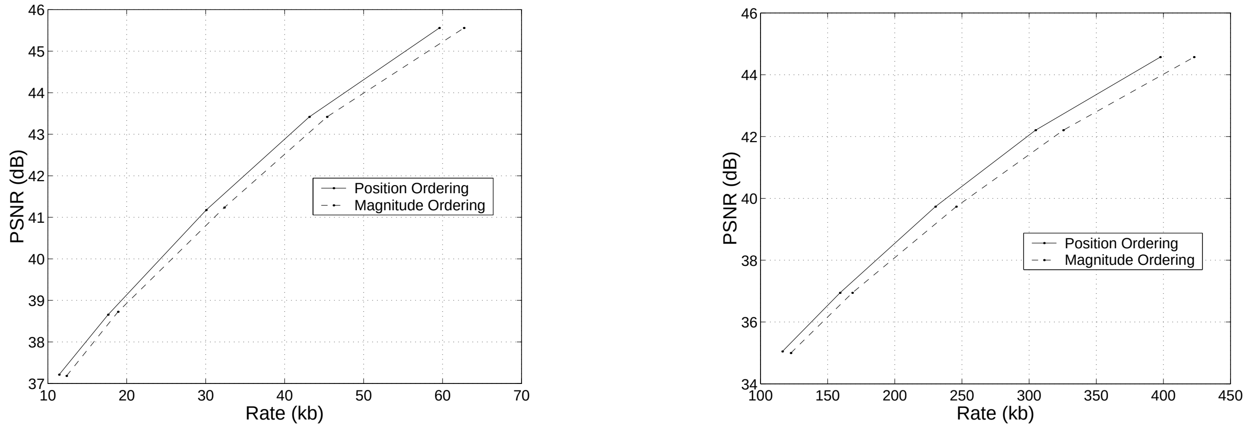
(b) Mobile CIF, fr.40.

Fig. 3. DFD Rate/Distortion curves comparing the optimal non-uniform quantizer (Opt. NULQ) with the close-to-optimal non-uniform mid quantizer (Mid. NULQ), which results from the high-resolution and constant atom coding cost assumptions made in simplifying (37) in (38). The fact that both curves are very close to each other demonstrates the relevance of these simplifications.

and the search for a new atom is only performed in the block with highest energy level [7]. Three uniform and one non-uniform quantizers are compared. All uniform quantizer stepsizes have been selected according to the simplified Equation (34), i.e. $\Omega/\Theta = 0.66\,\sqrt{\beta'^{-1}(\Theta)}$. The notation UQ refers to uniform quantizers for which the quantization error re-injection is neglected ($\beta'(\Theta) = 1$). UQ, IS is the approach used in [8] while UQ conforms to the stopping criterion expressed in Section II. So, the difference between them lies in the fact that UQ, IS stops the search as soon as an atom smaller than the threshold Θ has been encountered in one of the search subspaces, while UQ continues the search until no atoms larger than Θ can be found in any subspace. As shown in Fig. 2, UQ outperforms UQ, IS. The uniform quantizer denoted by ULQ takes into account the quantization error re-injection into the MP loop and sets $\beta'(\Theta)$ according to (20). The last quantizer, $NULQ$, is the non-uniform mid-quantizer designed from the simplified Equation (38). As expected, $NULQ$ outperforms all uniform quantizers. It is interesting to mention that, for both ULQ and $NULQ$, the β' or β'-sequence parameters are initialized to one. Their estimation is then refined according to the ISE approach described in Section IV. In practice, a single iteration is enough to achieve convergence.

Fig. 3 compares two non-uniform quantizers. The first one (Opti. NULQ) is optimal in the sense that it solves (37) and that it uses the reconstruction offset δ defined in (10). The second one (Mid. NULQ) is the mid-quantizer derived from (38). R/D curves are very close, demonstrating the validity of the high-resolution and constant atom coding cost assumptions, made in simplifying (37) in (38).

For DFD coding, Fig. 4 compares the scheme which encodes the positions differentially and the modulus independently [1], [8], with a scheme which orders the quantized atoms in decreasing order of magnitude and encodes the modulus differentially. For both schemes, uniform quantization is used and atoms are selected until their modulus achieves a given stopping threshold.



(a) Foreman QCIF, fr.60.

(b) Mobile CIF, fr.40.

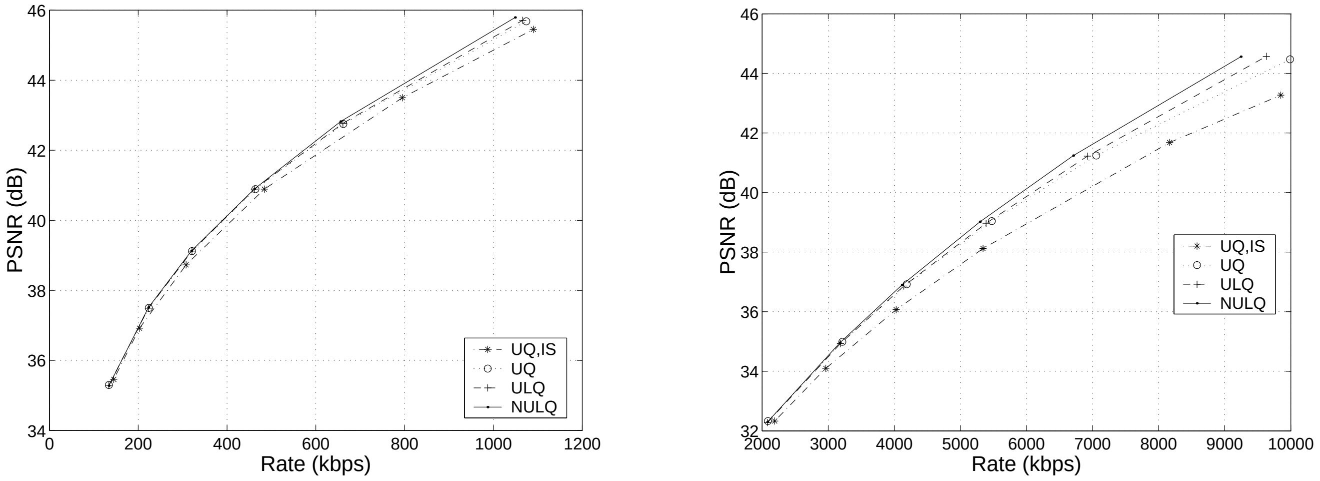
Fig. 4. DFD Rate/Distortion curves comparing the fixed length position encoding with the differential position encoding strategy. It motivates our study as it shows that when combined with efficient differential atom positions encoding schemes independent modulus quantization performs better than independent position encoding combined with differential modulus encoding.

For the scheme that represents the quantized modulus differentially, entropy provides a lower bound to the modulus encoding cost, and the quantizer stepsize that achieves the best R/D trade-off is selected using exhaustive search. As atoms are ordered in decreasing order of magnitude there is little correlation between successively encoded positions, so that the positions can be encoded independently with a fixed number of bits. By splitting the frame into blocks and varying the size of the blocks, one can trade-off the number of bits devoted to position coding and the differential modulus coding efficiency. The dashed R/D curves in Fig. 4 show the best performance obtained from an exhaustive search among block sizes and quantizer stepsizes. Although refined encoding schemes could be developed, we believe that the dashed curve provides a reasonable bound to the performance of the differential modulus encoding approach. The solid lines correspond to the framework studied in this paper and are obtained using the optimal uniform quantizer derived from (34) with $\beta\,'(\Theta) = 1$. As can be observed from the graphs, the solid line performs consistently better than the dashed line. We conclude that the differential encoding gain obtained on the positions compensates for the cost of independent modulus quantization. We can reasonably extend the conclusion to other applications than video coding. Indeed, as useful MP representations are sparse, they devote a large portion of the bit-budget to atom indices. As a consequence, a small entropy gain on the index representation becomes rapidly significant in comparison with the independent modulus description cost, motivating the study envisioned in this paper.

B. Video sequence atoms selection and quantization

In this section, we quantify the benefit of our quantizers when encoding entire video sequences, confirming the relevance of deriving the β' or β'-sequence parameters from the previously encoded frame.

Fig. 5 confirms that trends observed for isolated DFDs in Fig. 2 also apply to entire video sequences. Four mid-quantizers are compared. Both UQ, IS and UQ use the same uniform quantizers and neglect the quantization error re-injection, i.e.



(a) Foreman QCIF.

(b) Mobile CIF.

Fig. 5. Rate/Distortion curves for video sequences encoded at 30 fps. As in Fig. 2 each curve refers to a uniform or non-uniform mid-quantizer. UQ,IS and UQ are uniform quantizer defined by $\Omega/\Theta = 0.66$. UQ,IS corresponds to [10]. UQ goes on the search until no atom smaller than the threshold can be found in any part of the image. ULQ is the uniform quantizer defined by (34) with $\beta \neq 1$ while NULQ is the non-uniform mid-quantizer defined by (38). Both estimate the β' or β'-sequence parameters from the previous frame.

$\beta'(\Theta) = 1$ in (34). UQ, IS corresponds to the results obtained in [8] while UQ selects the atoms until no atom larger than the threshold can be found on the frame. ULQ and $NULQ$ take the re-injection into account and are defined by (34) and (38) respectively. For the uniform quantizer, the β' parameter is estimated from the previous frame. For the non-uniform quantizer, the β'-sequence has been estimated using the ISE method described in Section IV-B.2. As can be observed in the plots of Fig. 5, at low bitrates little improvement is to be expected from a sophisticated quantizer design. Most of the bits are indeed devoted to motion vector coding, and few atoms are encoded. On the contrary, as the bitrate increases, the non-uniform quantizer significantly outperforms [8]. The improvement depends on the sequence and lies between 0.5 dB and nearly 2 dB at high bitrates. In addition, for the non-uniform quantizer, we have observed that the two approaches proposed to estimate the β'-sequence, i.e. the ISE and RSC methods, perform equally well. Moreover, our experiments have shown that using RC-RSC with constant χ-sequence instead of the truncated one, i.e. the τ-sequence, does not significantly affect the performance. We have measured the coding loss in using RC-RSC as compared to RSC to tend to zero at low and high bitrates, and to be below 0.1 dB in between. This can be understood by the fact that the two sequences only differ on high order quantizer bins, i.e. on bins that contain few atoms. We conclude that it is relevant to define the non-uniform i^{th} quantizer stepsize as $\Omega_i = 0.66\ \chi_i\ \Theta$, i.e. proportional to the stopping threshold Θ, which makes it convenient to use in practice, as it virtually removes the non-uniform quantizer design complexity. For interested readers, Appendix D compares the performance of the MP video coder used in this paper with MPEG4 and MPEG4 AVC codecs. It demonstrates that, with equivalent motion estimation and compensation strategies, the MP expansion performs better than an approach based on orthogonal transforms.

VI. CONCLUSIONS

We have presented an R/D analysis of the MP expansion and quantization for systems that transmit atoms in random order of magnitude. This is certainly the case for video coding applications, where a significant entropy gain is obtained by differential description of the atom positions. Firstly, our study shows that, at optimality, atoms are selected up to a quality-dependent threshold. Another main result of the paper is that due to the re-injection of the quantization error during the MP expansion, the R/D optimal quantizer is non-uniform. A novel method is proposed to model the re-injection, and to design the optimal quantizer.

Our results rely on the observation that an MP expansion selects the atoms mostly in decreasing order of magnitude. Our developments do not assume a strictly decreasing atom selection order, and still hold when this property is valid on average, i.e. most of the time. To simplify the quantizer designs, the high-resolution assumption has been made, and we have assumed a near-constant expected atom coding cost. The relevance of these assumptions has been carefully analyzed. Moreover, to study re-injection, we have noted that each atom only captures a small fraction of the residual signal energy, and that quantization error is not the dominant part of the MP expansion reconstruction error.

When restricted to uniform quantizers, the analytical design closely matches the conclusions derived empirically in [8]. In particular, it validates the characterization of the quantizer by the ratio of its stepsize and its dead-zone. However, even for uniform quantization, the better understanding of the relation between the atom selection and quantization results in significant improvements of the results obtained in [8]. The R/D optimal non-uniform quantizer provides the largest improvements, especially at high bitrates, with gains larger than one dB compared to [8]. Moreover, at the end, each stepsize of the non-uniform quantizer can be defined proportionally to the stopping threshold. Using the observation that the residual signal energy decreases as a power of the atom selection stopping threshold, we have shown that the sequence of proportionality factors is signal independent, and can be pre-computed, which virtually removes the non-uniform quantizer design complexity.

In conclusion, to achieve R-D optimality, given a quality-dependent stopping threshold Θ, we recommend to expand the signal until all the atoms larger than Θ have been selected, and to use the non-uniform quantizer whose i^{th} quantizer stepsize is defined as $\Omega_i = 0.66\ \chi_i\ \Theta$. This formulation is very convenient to use in practice as the χ-sequence has appeared to be independent of the coded video sequence.

APPENDIX A

β-FACTOR ESTIMATION FOR UNIFORM QUANTIZATION

In this section, we demonstrate that the factor $\beta(N_\Theta)$ defined in equation (17) can be estimated as a function of the ratio between the initial residual energy, and the residual energy after N_Θ atoms have been selected.

To estimate $\beta(N_\Theta)$, we note that by definition, all γ_i are much smaller than one, and we approximate them by their mean value, denoted by$\overline{\gamma}$. The factor $\beta(N_\Theta)$ becomes

$$\begin{aligned} \beta(N_\Theta) &= \frac{\sum_{i=0}^{N_\Theta-1}\left(\Pi_{j=i+1}^{N_\Theta-1}(1-\gamma_i)\right)}{N_\Theta} \\ &\approx \frac{\sum_{i=0}^{N_\Theta-1}(1-\Sigma_{j=i+1}^{N_\Theta-1}\gamma_i)}{N_\Theta} \\ &\approx (1-N_\Theta\overline{\gamma}/2) \approx \sqrt{(1-N_\Theta\overline{\gamma})} \end{aligned} \tag{55}$$

In (13), with $n = N_\Theta$, and by neglecting the quantization distortion, i.e.$\sum_{i=0}^{N_\Theta}\left(\Pi_{j=i+1}^{N_\Theta}(1-\gamma_j)\right) D_{M_\Theta}(\Omega)$, in comparison with the energy of the residual signal, i.e. $\Pi_{i=0}^{N_\Theta}(1-\gamma_i)||\mathcal{R}_0 f||^2$, we have

$$\begin{aligned} ||\mathcal{R}_{N_\Theta} f||^2 &\approx \Pi_{i=0}^{N_\Theta-1}(1-\gamma_i)||\mathcal{R}_0 f||^2 \\ ||\mathcal{R}_{N_\Theta} f||^2 &\approx (1-\sum_{i=0}^{N_\Theta-1}\gamma_i)||\mathcal{R}_0 f||^2 \\ ||\mathcal{R}_{N_\Theta} f||^2 &\approx (1-N_\Theta\overline{\gamma})||\mathcal{R}_0 f||^2 \end{aligned} \tag{56}$$

From (55) and (56), $\beta(N_\Theta)$ can be estimated in terms of the ratio between the initial and final MP residual energy, i.e.

$$\beta(N_\Theta) \approx \sqrt{\frac{||\mathcal{R}_{N_\Theta} f||^2}{||\mathcal{R}_0 f||^2}} \tag{57}$$

By neglecting the quantization error in (13), these developments assume that the signal expansion error is the dominant part of the reconstruction error. This is a reasonable assumption since an MP expansion is made of costly atoms, due to index coding. So, if most of the reconstruction error was due to quantization, it would be beneficial to remove a few atoms from the expansion in order to improve quantization. Our simulations have confirmed this intuition.

APPENDIX B

β-SEQUENCE ESTIMATION FOR NON-UNIFORM QUANTIZATION

This section approximates the factor $\beta_i(N_\Theta, n_i, n_{i+1})$ defined in equation (25) as a function of the residual energy once n_i, n_{i+1}, and N_Θ atoms have been selected.

In equation (25), $\beta_i(N_\Theta, n_i, n_{i+1})$ depends on the unknown parameters γ_j, i.e.

$$\beta_i(N_\Theta, n_i, n_{i+1}) = \frac{\sum_{k=n_{i+1}+1}^{n_i}\left(\Pi_{j=k+1}^{N_\Theta-1}(1-\gamma_j)\right)}{n_i - n_{i+1}} \tag{58}$$

By taking the factor $\left(\Pi_{j=n_i}^{N_\Theta-1}(1-\gamma_j)\right)$ out of the sum in (58), we have

$$\beta_i(N_\Theta n_i, n_{i+1}) = \left(\Pi_{j=n_i+1}^{N_\Theta-1}(1-\gamma_j)\right) \cdot \frac{\sum_{k=n_{i+1}+1}^{n_i}\left(\Pi_{j=k+1}^{n_i}(1-\gamma_j)\right)}{n_i - n_{i+1}} \tag{59}$$

or equivalently for small γ_j

$$\beta_i(N_\Theta, n_i, n_{i+1}) \approx (1-\sum_{j=n_i+1}^{N_\Theta-1}\gamma_j) \cdot \frac{\sum_{k=n_{i+1}+1}^{n_i}(1-\Sigma_{j=k+1}^{n_i}\gamma_j)}{n_i - n_{i+1}} \tag{60}$$

Developments similar to the ones in (55) and (56) give

$$(1-\sum_{j=n_i+1}^{N_\Theta-1}\gamma_j) \approx \frac{||\mathcal{R}_{N_\Theta} f||^2}{||\mathcal{R}_{n_i} f||^2} \tag{61}$$

and

$$\frac{\sum_{k=n_{i+1}+1}^{n_i}(1-\Sigma_{j=k+1}^{n_i}\gamma_j)}{n_i - n_{i+1}} \approx \sqrt{\frac{||\mathcal{R}_{n_i}f||^2}{||\mathcal{R}_{n_{i+1}}f||^2}} \tag{62}$$

So, $\beta_i(N_\Theta, n_i, n_{i+1})$ is estimated as

$$\beta_i(N_\Theta, n_i, n_{i+1}) \approx \frac{||\mathcal{R}_{N_\Theta}f||^2}{||\mathcal{R}_{n_i}f||^2} \cdot \sqrt{\frac{||\mathcal{R}_{n_i}f||^2}{||\mathcal{R}_{n_{i+1}}f||^2}} = \frac{||\mathcal{R}_{N_\Theta}f||^2}{\sqrt{||\mathcal{R}_{n_i}f||^2 \cdot ||\mathcal{R}_{n_{i+1}}f||^2}} \tag{63}$$

and is approximated in terms of the residual signal energies after respectively n_i, n_{i+1} and N_Θ MP iterations, with $N_\Theta > n_i > n_{i+1}$.

APPENDIX C

RESIDUAL ENERGY DECREASE AS A FUNCTION OF THE STOPPING THRESHOLD: STUDY OF THE ρ PARAMETER

As told in Section IV-B.3, the parameter ρ defined in (35) is fundamental to build the non-uniform quantizer based on (54). For this reason, in this section, we analyze Equation (35) carefully.

Given two stopping thresholds θ_1 and θ_2, and denoting E_{θ_i} the residual energy obtained when selecting all atoms larger than θ_i, Equation (35) tells that

$$\frac{E_{\theta_1}}{E_{\theta_2}} = \left(\frac{\theta_1}{\theta_2}\right)^\rho \tag{64}$$

which, in words, means that, in logarithmic scales, the relation between the residual energy E_Θ and the stopping threshold Θ is linear. This property has been observed in Figure 1 for MP video coding, and is further studied in Figure 6 for generic 1-D signals. Specifically, Figure 6 presents the residual energy E_Θ as a function of the stopping threshold Θ for random signals in R^M that are expanded on two different dictionaries. One dictionary is composed of maximally spaced points on the unit sphere, i.e. it minimizes the maximal distance between two dictionary functions [11], while the other dictionary is made of random vectors in R^M. For both cases, we observe that (64) is true, i.e. that the relation between E_Θ and Θ is linear in logarithmic scales. In Figure 6, we also observe that for both dictionaries, the slope of the linear part of the graph is close to 2, i.e. $\rho \approx 2$.

We now demonstrate that ρ is upper-bounded by 2. For any signal f in R^M, let α_i denote the modulus of the atom extracted at step $i \geq 0$ and $\mathcal{R}_i f$ denote the residual energy after i atoms have been extracted. Let γ_i denote the fraction of the residual signal energy captured by the MP expansion at step $i \geq 0$. We have thus $\gamma_i = \frac{\alpha_i^2}{||\mathcal{R}_i f||^2}$, with $||.||$ referring to the quadratic norm in R^M. Given the stopping thresholds $\theta_1 < \theta_2$ and assuming that atoms are selected in decreasing order of magnitude, which is true most of the time, there exists two iteration indices $i_1 > i_2$, so that $\alpha_{i_j} = \theta_j$ and $||\mathcal{R}_{i_j}f||^2 = E_{\theta_j}$ for $j = 1, 2$. By definition of the γ_i factors, $\gamma_{i_j} = \theta_j^2 / E_{\theta_j}$, and we have

$$\frac{E_{\theta_1}}{E_{\theta_2}} = \frac{\gamma_{i_2}}{\gamma_{i_1}} . \frac{\theta_1^2}{\theta_2^2},\ i_1 > i_2. \tag{65}$$

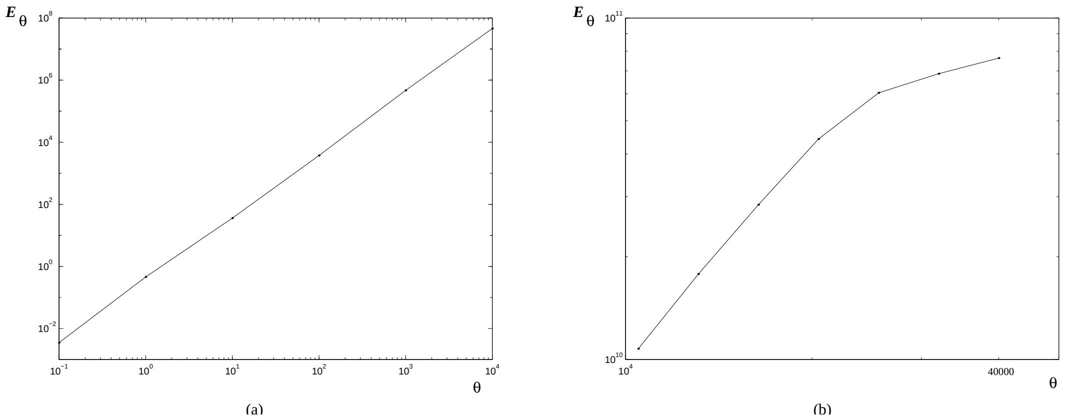
In [6], Mallat and al. have shown that the expected value of γ_i decreases with i, and tends to a constant value when i increases to infinity. As a consequence, in (65), for any $i_2 < i_1$ that tends to infinity, $\gamma_{i_1} \approx \gamma_{i_2}$ and we have

$$\frac{E_{\theta_1}}{E_{\theta_2}} \approx \left(\frac{\theta_1}{\theta_2}\right)^2 \tag{66}$$

which is equivalent to (64) with $\rho = 2$. In [6], it has been observed that the expected value for γ_i rapidly converge to a constant value for noisy signals. It explains why the curves in Figure 6 rapidly become linear, and why $\rho = 2$ in Figure 6.

More generally, introducing (64) in (65) gives

$$\frac{\gamma_{i_1}}{\gamma_{i_2}} = \left(\frac{\theta_1}{\theta_2}\right)^{2-\rho} \tag{67}$$



(a)

(b)

Fig. 6. Residual energy E_Θ appears to be a linear function of the stopping threshold Θ. Logarithmic scales are used for both axes, and curves average the energy obtained when expanding 10 random signals in R^M on two different dictionaries. On the left: signals are in R^5, and the dictionary is composed of 128 maximally spaced points on the unit sphere [11]. On the right: signals are in R^{500}, and the dictionary is composed of 5000 random vectors in R^{500}.

To ensure that the expected value of γ_i decreases with i, the expected value of ρ has to be lower or equal to 2. Indeed, with $\rho \leq 2$, (67) ensures that for any $\theta_1 < \theta_2$, or equivalently any $i_1 > i_2$, we have $\gamma_{i_1} < \gamma_{i_2}$. For video coding, we have observed empirically in Figure 1 that $\rho \approx 1.5$, which shows that $\gamma_{i_1}/\gamma_{i_2} \approx \sqrt{\theta_1/\theta_2}$. The study of this property remains to be done, but is beyond the scope of the paper.

Appendix D

MP vs MPEG4 video coding

Figure 7 compares the video coder considered in this paper, and defined in [1], [7], with MPEG4 and MPEG4-AVC ISO/IEC standards. As the motion estimation and compensation engine of the MP coder in [1], [7] is very similar to the motion prediction engine defined by MPEG4 (half-pel, 8x8 blocks), the comparison between MPEG4 and our MP coder is relevant. In Figure 7, we observe that the MP approach significantly outperforms MPEG4. This is in accordance with the conclusions drawn in [7] and [8] for low bitrates. The improvement achieved by our proposed non-uniform quantization strategy makes it possible to extend this conclusion to the entire range of bitrates.

For completeness, we also present the rate/distortion characteristic obtained with the complex but high performance MPEG4-AVC coder. The comparison with the MP approach is questionable in the sense that MPEG4-AVC is based on a fine, adaptive, and rate/distortion optimized motion compensation. Specifically, MPEG4-AVC performs $(1/8)^{th}$ of pel motion compensation on blocks of adaptive sizes, and optimizes the choice among a large set of prediction modes according to rate/distortion criteria. On the contrary, MP is simply based on half-pel motion compensation of 8x8 blocks.

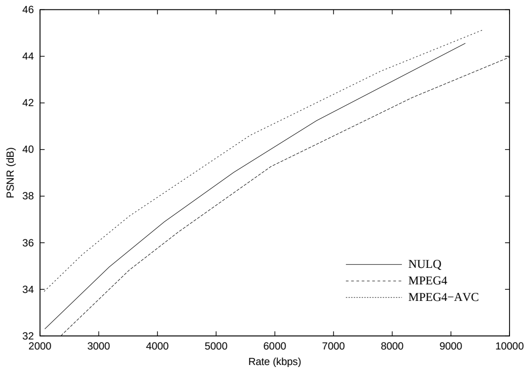


Fig. 7. Rate/Distortion curves obtained when encoding Mobile, CIF, 30fps. The NULQ curve refers to MP coding with the proposed non-uniform quantizer defined by (38). MPEG4 refers to the MPEG4 video coding standard, while MPEG4-AVC refers to the advanced video coding standard currently defined by ISO. NULQ and MPEG4 use similar motion estimation and compensation engines. On the contrary, MPEG4-AVC is based on much finer motion compensation and rate/distortion optimization rules to improve the prediction stage of the video coder.

REFERENCES

[1] O. Al-Shaykh, E. Miloslavsky, T. Nomura, R. Neff, and A. Zakhor. Video compression using matching pursuits. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1):113–143, February 1999.

[2] P. Frossard and P. Vandergheynst. A posteriori quantized matching pursuit. In *Proceedings of the IEEE Data Compression Conference*, Snowbird, UT, March 2001.

[3] V.K. Goyal, M. Vetterli, and T.T. Nguyen. Quantized overcomplete expansions in R^N: analysis, synthesis, and algorithms. *IEEE Transactions on Information Theory*, 44(1):16–31, January 1998.

[4] R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):1–63, October 1998.

[5] S. Hein and A. Zakhor. Reconstruction of oversampled bandlimited signals from sigma delta encoded binary sequences. *IEEE Transactions on Signal Processing*, 42(4):799–811, March 1994.

[6] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.

[7] R. Neff and A. Zakhor. Very low bit rate video coding based on matching pursuits. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):158–171, February 1997.

[8] R. Neff and A. Zakhor. Modulus quantization for matching pursuit video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):895–912, September 2000.

[9] R. Neff and A. Zakhor. Matching-pursuit video coding - Part I: Dictionary approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(1):13–26, January 2002.

[10] R. Neff and A. Zakhor. Matching-pursuit video coding - Part II: Operational models for rate and distortion. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(1):27–39, January 2002.

[11] N.J.A. Sloane, R.H. Hardin, and W.D. Smith. Spherical codes: a library of best ways to place n points on a sphere in d dimensions so as to maximize the minimal distance between them. *Available online at http://www.research.att.com/ njas/packings*.

[12] G.J. Sullivan. Efficient scalar quantization of exponential and laplacian random variables. *IEEE Transactions on Information Theory*, 42(5):1365–1374, September 1996.