Mutual Information-Based Context Quantization

Marco Cagnazzo[a], Marc Antonini[b], Michel Barlaud[b]

[a] *TELECOM-ParisTech, 46 rue Barrault, 75634 Paris (FRANCE)*
[b] *I3S Laboratory, 2000 route des Lucioles, 06903 Sophia Antipolis (FRANCE)*

Abstract

Context-based lossless coding suffers in many cases from the so-called *context dilution* problem, which arises when, in order to model high-order statistic dependencies among data, a large number of contexts is used. In this case the learning process cannot be fed with enough data, and so the probability estimation is not reliable. To avoid this problem, state-of-the-art algorithms for lossless image coding resort to context quantization (CQ) into a few conditioning states, whose statistics are easier to estimate in a reliable way. It has been early recognized that in order to achieve the best compression ratio, contexts have to be grouped according to a maximal mutual information criterion. This leads to quantization algorithms which are able to determine a local minimum of the coding cost in the general case, and even the global minimum in the case of binary-valued input.

This paper surveys the CQ problem and provides a detailed analytical formulation of it, allowing to shed light on some details of the optimization process. As a consequence we find that state-of-the-art algorithms have a suboptimal step. The proposed approach allows a steeper path toward the cost function minimum. Moreover, some sufficient conditions are found, that allow to find a globally optimal solution even when the input alphabet is not binary. Even though the paper mainly focuses on the theoretical aspects of CQ, a number of experiments to validate the proposed method have been performed (for the special case of segmentation map lossless coding), and encouraging results have been recorded.

Key words: Lossless coding; context dilution; context quantization.

Email addresses: cagnazzo@telecom-paristech.fr (Marco Cagnazzo), am@i3s.unice.fr (Marc Antonini), barlaud@i3s.unice.fr (Michel Barlaud)

1. Introduction

Entropy coding is a key component for many visual data compression techniques, be them devoted to lossless image coding (JPEG-LS [1]), to lossy image coding (JPEG [2], JPEG2000 [3]) or to video coding (MPEG-1, 2 or 4 [4, 5, 6], H.264 [7]). It aims at representing the realization $z^n = (z_1, z_2, \ldots, z_n)$ of a random vector Z^n with the minimum possible number of bits, that is $-\log p(z^n)$[1]. Averaging over the probability of each vector z^n, the lower bound of the coding cost for a message generated by the random process Z^n is the entropy $H(Z^n)$. However, in order to approach this bound it is necessary to dispose of the joint probability of the whole message. A practical way to do this is suggested by the entropy chain rule [8],

$$H(Z_1, Z_2, \ldots, Z_n) = \sum_{i=1}^{n} H(Z_i|Z^{i-1}).$$

This equation assures that a sequential coding of $Z_i|Z^{i-1}$ achieves the same rate of the joint coding of Z^n. However, this approach is useful only if the conditional probabilities $p(z_i|z^{i-1})$ are easier to compute than the overall probability $p(z^n)$. This is true for example when the current sample depends only on a fixed and finite subset of the previous ones, called the causal pattern.

Let Y be a generic input symbol, $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ be the input alphabet, and N be the number of samples constituting a context X. The set of all contexts $\mathcal{X} = \{x_1, x_2, \ldots\}$ is $\mathcal{Y}^N$. We assume that the conditional probability of the current symbol Y with respect to the whole past sequence coincides with the conditional probability with respect to a suitable causal template X. In this case, an entropy coder disposing of an estimation of $p(Y|X = x)$ for any x, let it be $\hat{p}(Y|x)$, can encode Y with $-\log \hat{p}(Y|x)$ bits. The estimated probabilities can be pre-computed off-line through a training set, or evaluated on the fly on the basis of the previously encoded symbols, for example by using their relative frequencies. This is the adaptive model, while the former is the static one. It can be shown [9] that the adaptive model performs only slightly worse than the best possible static model, which in turns requires a two-passes coding. For this reason, the adaptive model is widely employed.

[1]Throughout this paper, the logarithm is in base 2 and the symbol $p(\cdot)$ is used for the probability mass functions; the corresponding random vector is identified by its argument.

However, when the input symbol alphabet cardinality M or the causal template size N are large, the conditional probability estimation can face a severe *context dilution*: having too many contexts makes it difficult or practically impossible to estimate and update the conditional probabilities of symbols during the encoding process. To avoid this problem, state-of-the-art algorithms for lossless image compression like CALIC [10] and the arithmetic encoder in EBCOT [11], as well as popular algorithms proposed in the scientific literature, resort to *context quantization* (CQ). CQ consists in grouping contexts into a relatively small number of conditioning states $c_1, c_2, \ldots, c_F$. Each state, or context class, c_i is made up of one or more contexts x, according to the quantization function which associates a context to a class label, $f : x \in \mathcal{X} \rightarrow f(x) \in L = \{1, 2, \ldots, F\}$. Thus, the class c_i is made up of all the contexts $x \in \mathcal{X}$ such that $f(x) = i$. Now we only need the F probability mass functions (pmf's) $p(Y|f(x) = i)$ (also referred to as $p(Y|c_i)$) instead of the M^N pmf's $p(Y|x)$. Of course, the counterpart is that the CQ reduces the mutual information (MI) between the current symbol and the conditioning state. It has been shown that this MI loss affects directly the coding performances, as it is translated into an equal coding rate increase. We come back to this point in the following (see Section 3), but we remark here the importance of designing the best possible context quantizer.

In this paper we first review the state of the art for CQ (Section 2), with particular focus on the case of a fixed number F of conditioning states. We also expose the reasons and the alternatives of this choice. In Section 3 we perform a detailed analysis of this problem, providing a unified framework. A first original contribution of the present work, presented in section 4, is that our analysis sheds light on some fine details of the CQ process with a given F, allowing a complete control on it. A new result is that the iterative algorithms presented in the literature do not completely take into account all the effects of modifying an assigned context classification, and so they do not perform the most effective operations at each step. Our framework allows to find the algorithm achieving the largest improvement of the cost function at each step. Moreover, this algorithm proves to be superior to the classical ones even from an experimental point of view, as we will show later. In Section 5 we find some sufficient conditions which allow to find the globally optimal classification even when the input alphabet has more than two symbols (even if we consider cases where the influence of the context is in some way binary). Section 6 provides experimental compression performances of the proposed CQ algorithm for the motion segmentation maps of a set of

test video sequences. These results confirm the improvements achieved by our method. Finally Section 7 draws conclusions and ends the paper.

2. Background and Prior Work

Context quantization is a major issue in lossless coding of images. However we favor a more general approach, and consider generic bi-dimensional arrays. This means in particular that the statistic dependencies among samples can be other than simply linear, as in the case of motion segmentation maps described by parametric contours [12] which we consider in the experimental section. In order to keep generality we do not make any assumption on the kind of dependence among input samples.

With respect to lossless coding of bi-dimensional data, a seminal paper is the one by Weinberger *et al.* [13], where universal coding is introduced for this kink of data. In this paper, authors consider the problem of context-based coding of gray-level images. The construction of the contexts takes into account the so-called *model cost*. That is, a too complex model involves a more costly description and a less accurate parameter estimation which can cancel out the benefits of the attempt to account for high-order dependencies. This problem is also called *context dilution*. If too many contexts are considered, the estimation of $p(Y|x)$ can be unreliable or difficult. The general solution is *context quantization* or *classification*. This means that many contexts are put together to form a context class c_i which is the preimage of i under the classification function f: $c_i = \{x \in \mathcal{X} : f(x) = i\}$. The f function can be seen as a quantization function, since it groups many contexts to the same class. Since the number of classes is generally small, the context dilution problem is avoided, even though this result comes at the cost of an increase of the achievable coding rate. The solution proposed in [13] is to adaptively build a causal pattern for CQ. This is obtained by a tree structure such that only contexts influencing the current symbol *exactly* in the same way are merged together. This method relies on the linear dependencies typical of image data, so it cannot be directly applied to more general cases. Another difference between [13] and our method is that in our case the number of classes F is given, while in [13] it is determined by an adaptive algorithm which however makes several assumptions based on the characteristics of gray-scale images. Our choice of studying the CQ for a given number of classes fits with the cases where constraints on complexity and memory needs exist, and indeed, many papers on CQ assume the same framework. In these works,

it was early recognized that the optimal CQ given F is done by maximizing the mutual information $I(Y, f(X))$, or equivalently by minimizing the conditional entropy $H(Y|f(X))$, since the latter quantity coincides with the increase of the minimum coding rate due to CQ.

One of the first works where these concepts are employed is [14] by Wu, where the author deals with the problem of optimal CQ for lossless grey-scale image coding. As in this case the input signal has a large alphabet, the context dilution problem is important even with a very small template. However, although it is recognized that the optimal CQ should maximize the mutual information, this is done in an *ad hoc* way, since the context is represented only by a so called error discriminant (taking into account how much pixels in the template differ from a linear prediction of the current sample). Instead of considering all the possible values of error discriminant, these are quantized on a fixed number of levels with a maximal mutual information criterion. This strategy depends on the prediction algorithm, which in turns relies on linear dependencies of data.

In [15] it is recognized once more that the CQ with a given number of classes should be driven by a conditional entropy minimization, but authors proposed instead the minimization of the prediction error energy since this suboptimal choice allow to develop the mathematical analysis of the problem, with a solution claimed to be generally quite near to the optimum.

The problem of optimal CQ with a given F is faced by Green *et al.* in [16] for bi-level images. In this paper it is recognized that the correct cost function is the mutual information between the current symbol and the context class, and that it can be expressed as a weighted mean of relative entropies (or Kullback-Leibler divergences) between conditioned pmf's. As reported in [16], this cost function was firstly introduced in the case of text compression by Bookstein *et al.* [17]. The paper by Green *et al.* analyzes only the case of a binary input, and it introduces an algorithm based on dynamic programming, with a dramatic complexity reduction.

The paper [18] and even more relevantly its extended version [19] by Wu *et al.* are however the first ones in which the optimal CQ with a given F is conveniently analyzed and a generic analytical solution is proposed. The optimal quantization problem is recognized as a special case of vector quantization, with the Kullback-Leibler divergence to be used as distortion measure. As a consequence, a GLA-like algorithm is proposed to find the optimal CQ, which must minimize the conditional entropy between the current symbol and the quantized context. This approach is then called minimum condi-

tional entropy context quantization (MCECQ). A first original contribution of our work is an analysis of the mathematical background of MCECQ, which allows to discover a suboptimal step and then to propose an improved algorithm for CQ (this is a second contribution). As already seen in [16], the paper [18] recognizes that if Y has a binary distribution, then the optimal CQ problem is brought back to a 1D minimization problem, which can be *exactly* solved by means of dynamic programming techniques. As third contribution, we show how to extend this approach to M-valued symbols (with $M > 2$), even if the distributions that we consider have essentially a binary behavior. All relevant subsequent papers adopt approaches similar to those of [16] or [18]. This is the case for example of [20] by Chen. This paper develops an analysis of the CQ problem for the lossless coding of WT coefficients and ends up with an algorithm equivalent to MCECQ.

The paper [21] by Forchhammer *et al.* takes [18] as a starting point to develop an optimal CQ technique for binary symbols. Here the accent is on the binarization of input data samples, which allows to use dynamic programming to find the globally optimal CQ. Moreover the authors propose the minimum code length context quantization (MCLCQ), which is an alternative criterion to MCECQ and takes into account the specific probability estimator. As a consequence, the optimal solution depends on the parameters of this estimator. In [21] and [22] experimental results showed that MCECQ and MCLCQ allow remarkable coding rate reductions with respect to state-of-the-art techniques, when used to code MPEG4 α-plane video sequences. MCLCQ is reconsidered in [23], where Forchhammer and Wu approach the important problem of choosing the optimal number of classes. The target is to minimize the expected adaptive coding length with respect both to the classification function and the number of classes.

In [24], Liu and Karam use optimal CQ to improve the contexts for the JPEG2000 arithmetic coder. In this case the input symbols are binary, and so dynamic programming can be used to find the globally optimal CQ. The number of classes is fixed *a priori*, but a study on the impact of this choice on the overall performance is reported. The approach of this paper is interesting and some other research groups have followed it [25].

We conclude this survey by considering the problem of CQ within coding standards. In the JPEG-LS standard (based on the LOCO-I algorithm [1]) it is recognized that optimal CQ should be achieved by mutual information maximization, but the proposed solution trades off between performance and complexity, and relies on a correlation-based model of the input signal. In

MPEG-4 some tools are provided for the lossless coding of binary maps [26]. The proposed algorithm uses a quite large context and an arithmetic encoder, but the problem of optimal CQ is not considered. In the CABAC coder [27] of the H.264 standard, the context classification is performed on *ad hoc* basis rather than following an optimization criterion.

In conclusion, several papers in the scientific literature are strictly related to the CQ issue. Like in our case, in several papers no *a priori* assumption about data (like linear dependencies) is made, [16, 18, 19, 24], so we naturally refer to these ones, even if interesting ideas are found in the others (like [1, 14, 27]) as well.

3. The context quantization problem

In this section we review the problem of CQ and the derivation of the optimal criterion. The final result had been shown in [19], and used in [20] and in [24], but we come back to it in order to give a unified framework and to better highlight the original contributions of our paper, presented in sections 4, 5 and 6.

We start with the notation. Let Y be the current symbol of a random process having a discrete[2] alphabet $\mathcal{Y}$. We assume without loosing generality $\mathcal{Y} = \{1, 2, \ldots, M\}$. X is the context, formed by N causal neighbors of the current symbol[3], *i.e.* by N already encoded symbols. The alphabet of X is $\mathcal{X} = \mathcal{Y}^N$, thus the number of possible contexts is $|\mathcal{X}| = M^N$. This number can grow up very large, and we risk to incur in the *context dilution* problem: the estimation of $p(Y|x)$, necessary for the encoder to achieve the minimum coding rate, is unreliable because there are too many contexts and usually not enough data. In conclusion we want to encode Y using previously encoded symbols information, attaining in this way a coding cost bounded by $H(Y|X)$ bits per symbol, but to reach this target we should estimate *on-the-fly* the probability distribution $p(Y|X)$, which is usually impossible or at least unreliable, because of the context dilution. So we consider a classification function $f : x \in \mathcal{X} \to i = f(x) \in L = \{1, 2, \ldots, F\}$.

The classification function f maps each context in a class label i inducing a partition $\mathcal{C}$ of $\mathcal{X}$ into F classes: $\mathcal{C} = \{c_1, c_2, \ldots, c_F\}$. We use these classes

[2]The hypothesis that Y is discrete is not necessary, but it simplifies the treatment.

[3]It is not necessary to impose that the context is formed by previously encoded symbols, but we keep this hypothesis for simplicity.

as conditioning information to encode Y, because with a suitable choice of F the estimation of the pmf's $p(Y|c)$ is reliable enough. In the following we consider the problem of optimizing the CQ for a given F. This approach is simpler than a joint optimization of the number of classes and of their composition, but it is popular [16, 18, 19, 20, 24] and sometimes necessary, as in the case of implementation constraints over memory or complexity. On the other hand, a joint optimization has the potential of achieving the best performance. However, a common approach [20, 24] consists in developing an optimization for a given F and then in testing different values of it.

In conclusion, we consider a framework where we can attain a new entropy bound on the coding cost, that is $H(Y|f(X))$. In order to analyze the relationship between the new and the old coding rate bounds, we observe that Y and $f(X)$ are conditionally independent given X: indeed $f(X)|X$ is deterministic and hence independent from $Y|X$. In conclusion, we have:

$$\begin{aligned} I(Y;X|f(X)) &= H(Y|f(X)) - H(Y|X,f(X)) \\ &= H(Y|f(X)) - H(Y|X) \end{aligned} \tag{1}$$

where in the definition of conditional mutual information [8] we used the conditional independence of Y and $f(X)$ given X. This equation tells us that the rate we can achieve with CQ, $H(Y|f(X))$ is greater than (or equal to) the rate we could achieve without classification, $H(Y|X)$, and that the increase is the mutual information $I(Y;X|f(X))$. We indicate this quantity with $\mathcal{L}(f)$ or $\mathcal{L}(\mathcal{C})$ because it is the *loss* associated to the classification function f or equivalently to the partition $\mathcal{C}$. In conclusion, the classification function should be chosen so that this loss is minimized for a given number of context classes. Now we derive a useful formulation of $\mathcal{L}(f)$. Let us use the shorthand notations $p(y,i)$, $p(y|i)$ for $\Pr(Y=y, f(X)=i)$, $\Pr(Y=y|f(X)=i)$, and $p(x,y)$, $p(x|y)$ for $\Pr(Y=y, X=x)$, $\Pr(Y=y|X=x)$.

The probability of a context class c is $p(c) = \sum_{x\in c} p(x)$. We consider the conditional entropy $H(Y|f(X))$: since $p(y,i) = \sum_{x\in c_i} p(y,x)$, we write it as:

$$H(Y|f(X)) = -\sum_{i\in L}\sum_{y\in\mathcal{Y}}\sum_{x\in c_i} p(y,x)\log p(y|i). \tag{2}$$

On the other hand, we can express $H(Y|X)$ decomposing the set $\mathcal{X}$ into the context classes:

$$H(Y|X) = -\sum_{i\in L}\sum_{x\in c_i}\sum_{y\in\mathcal{Y}} p(y,x)\log p(y|x). \tag{3}$$

So, by combining (1), (2) and (3), the mutual information loss becomes:

$$\mathcal{L}(f) = \sum_{i \in L} \sum_{x \in c_i} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{p(y|i)} \tag{4}$$

We can read the internal sum in (4) as the relative entropy between $p(Y|x)$, the distribution of the current symbol given a realization x of the context X, and $p(Y|i)$, the distribution of Y given a realization $i = f(x)$ of the conditioning class. If c is the class associated with the label i, *i.e.* if c is the preimage of i, we will also indicate $p(Y|i)$ as $p(Y|c)$. The mutual information loss can finally be written as:

$$\begin{aligned} \mathcal{L}(f) &= \sum_{i \in L} \sum_{x \in c_i} p(x) D\left(p(Y|x) \;||\; p(Y|i) \right) \\ &= \sum_{x \in \mathcal{X}} p(x) D\left(p(Y|x) \;||\; p(Y|f(x)) \right), \end{aligned} \tag{5}$$

where we used the notation $D\left(\cdot \;||\; \cdot \right)$ to indicate the relative entropy [8]. If we define $d(c_1, c_2)$ as

$$d(c_1, c_2) = D\left(p(Y|c_1) \;||\; p(Y|c_2) \right), \tag{6}$$

we can rewrite (5) as

$$\mathcal{L}(f) = \sum_{x \in \mathcal{X}} p(x) d\left(x, c_{f(x)}\right), \tag{7}$$

where with a little abuse of notation, in the expression of $d(\cdot, \cdot)$ we indicate with x the class $\{x\}$, and $c_{f(x)}$ is the class containing the context x. We can read (7) in this way: the global loss of mutual information due to CQ f, is the average loss that we incur in by substituting the actual context x with its quantized version $c_{f(x)}$. The loss for a single context is a measure of how much its influence on the current symbol is well represented by the *class* influence. If this influence is strictly the same (i.e. $p(Y|x) = p(Y|f(x))$) there is no loss; otherwise the loss is the relative entropy between the two probability distributions.

In order to establish a link with prior works, we observe that our cost function, the mutual information loss (7), is equivalent to the loss function definition in [16], and to equation (1) in [18], equation (2) in [19] or equation (6) in [20]. Moreover, it is slightly more general than equation (7) in [24], which only allows to group together contexts with consecutive indexes.

4. Context quantization properties and the MINIMA algorithm

In previous papers, the CQ was performed by using iterative algorithms looking for a local minimum of the cost function by starting from an arbitrary partition $\mathcal{C}_0$ and then moving contexts from one class to another. A remarkable exception is the case of bi-level images (*i.e.* data with a binary alphabet), where the global minimum of the cost function can be found via dynamic programming. We will come back to this case in Section 5; here we consider the general M-ary case, and we use the formulation given in the previous section to analyze the effect of these context displacements, and then to validate the approaches proposed in previous works. In this framework, we find that popular context displacement strategies do not perform necessarily the best move at each step, *i.e.* the one that maximally reduces the mutual information loss. Moreover we propose a "greedy" CQ algorithm, which at each step chooses the single context displacement assuring the best improvement of the cost function.

In this analysis we make intensive use of the relative entropy between the conditional probabilities functions $p(Y|c_1)$ and $p(Y|c_2)$, referred to as $d(c_1, c_2)$. If c_i is a single context, we will prefer the simpler notation $d(x, c)$ to the more correct one $d(\{x\}, c)$. With a little abuse, we will refer to this quantity as *distance* between x and c. We start by computing the effect of the insertion of a context x into a class c. It is intuitive that this operation reduces the relative entropy $d(x, c)$, since it amounts to injecting a term that is proportional to $p(Y|x)$ into the sum giving $p(Y|c)$.

Proposition 1. *Inserting a context into a class reduces the context distance from the class. Namely, if $x \notin c$ and $c' = c \cup \{x\}$, then*

$$d(x, c') = d(x, c) - \frac{p(c')}{p(x)} d(c', c) - \frac{p(c)}{p(x)} d(c, c'). \tag{8}$$

Proof. We observe that $p(y|c) = \sum_{x \in c} \frac{p(x)}{p(c)} p(y|x)$, and so:

$$p(y|c') = \frac{p(x)}{p(c')} p(y|x) + \frac{p(c)}{p(c')} p(y|c)$$

$$p(x)p(y|x) = p(c')p(y|c') - p(c)p(y|c). \tag{9}$$

Now we can write:

$$\begin{aligned}
p(x)d(x,c) =& p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|c)}\\
=& p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|c)}\frac{p(y|c')}{p(y|c')}\\
=& p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|c')} + \sum_{y\in\mathcal{Y}} p(x)p(y|x)\log\frac{p(y|c')}{p(y|c)}\\
=& p(x)d(x,c') + \sum_{y\in\mathcal{Y}}[p(c')p(y|c') - p(c)p(y|c)]\log\frac{p(y|c')}{p(y|c)} \qquad (10)\\
=& p(x)d(x,c') + p(c')d(c',c) + p(c)d(c,c').
\end{aligned}$$

Eq. (10) is obtained by using (9). □

The next proposition allows us to understand the effect on $\mathcal{L}(\mathcal{C})$ of moving a context from a class to another.

Proposition 2. *When a context x is moved from a class c_1 to a class c_2, the total mutual information loss varies of:*

$$\Delta = p(c_2)d(c_2,c_2') - p(c_1')d(c_1',c_1) + p(x)[d(x,c_2') - d(x,c_1)], \qquad (11)$$

where $c_1 = \{x_1, x_2, \ldots, x_N, x\}$, $c_2 = \{z_1, z_2, \ldots, z_M\}$, $c_1' = c_1 - \{x\}$ and $c_2' = c_2 \cup \{x\}$

Proof. Let $\mathcal{C} = \{c_1, c_2, c_3, \ldots c_F\}$ and $\mathcal{C}' = \{c_1', c_2', c_3, \ldots c_F\}$. We have $\Delta = \mathcal{L}(\mathcal{C}') - \mathcal{L}(\mathcal{C})$. If we indicate with $l(c)$ the contribution of class c to the total cost function, *i.e.* $l(c) = \sum_{z\in c} p(z)d(z,c)$, we have $\mathcal{L}(\mathcal{C}) = \sum_{i=1}^{F} l(c_i)$, and so:

$$\Delta = l(c_1') - l(c_1) + l(c_2') - l(c_2).$$

Moreover

$$d(x_i,c_1') - d(x_i,c_1) = \sum_{y\in\mathcal{Y}} p(y|x_i)\log\frac{p(y|x_i)}{p(y|c_1')}\frac{p(y|c_1)}{p(y|x_i)} = \sum_{y\in\mathcal{Y}} p(y|x_i)\log\frac{p(y|c_1)}{p(y|c_1')}$$

and

$$\begin{aligned} l(c_1') - l(c_1) &= \sum_{i=1}^{N} p(x_i) d(x_i, c_1') - \sum_{i=1}^{N} p(x_i) d(x_i, c_1) - p(x) d(x, c_1) \\ &= \sum_{y \in \mathcal{Y}} \sum_{i=1}^{N} p(x_i) p(y|x_i) \log \frac{p(y|c_1)}{p(y|c_1')} - p(x) d(x, c_1) \\ &= \sum_{y \in \mathcal{Y}} p(c_1') p(y|c_1') \log \frac{p(y|c_1)}{p(y|c_1')} - p(x) d(x, c_1) \\ &= -p(c_1') d(c_1', c_1) - p(x) d(x, c_1). \end{aligned} \tag{12}$$

Likewise,

$$l(c_2') - l(c_2) = p(c_2) d(c_2, c_2') + p(x) d(x, c_2'). \tag{13}$$

Adding (12) and (13), we get the thesis. □

Of course the Δ resulting from a context displacement can be either positive or negative, *i.e.*, the displacement can worsen or improve the CQ. This is because the effects on the classes c_1 and c_2 have opposite signs: $l(c_1') - l(c_1)$ is negative since removing a context from a class allows its pmf to be closer to the pmf's of remaining contexts[4], so that the loss of mutual information becomes smaller; for the symmetrical reason, adding a context to c_2 increases its contribution to $\mathcal{L}$. Thus, in order to tell whether a context displacement is suitable or not, we should consider both contributions, and verify that the gains surpass the losses. We remark that the results of propositions 1 and 2 are new; prior works, on the contrary, used directly (7) to derive an iterative algorithm for mutual information loss minimization. This algorithm is based on a property demonstrated in [19], and used in [20] and [24]. With the next proposition we provide a new demonstration of this property, which is useful in order to compare our algorithm with literature.

Proposition 3. *Let*

$$\begin{aligned} c_1 &= \{x_1, x_2, \ldots, x_N, x\} & c_2 &= \{z_1, z_2, \ldots, z_M\} \\ c_1' &= c_1 - \{x\} & c_2' &= c_2 \cup \{x\} \end{aligned}$$

[4]For example, if $c_1 = \{x, z\}$, then $c_1' = \{z\}$ and so $l(c_1') = 0$, because $p(y|c_1') = p(y|z)$.

$$\mathcal{C} = \{c_1, c_2, c_3, \dots c_F\} \qquad \mathcal{C}' = \{c_1', c_2', c_3, \dots c_F\}.$$

If $d(x, c_1) \geq d(x, c_2)$, the new context partition obtained by switching x from c_1 to c_2 has a smaller MI loss.

PROOF. By hypothesis, $d(x, c_1) \geq d(x, c_2)$. Combining with (8), we get

$$p(x)d(x, c_1) \geq p(x)d(x, c_2') + p(c_2')d(c_2', c_2) + p(c_2)d(c_2, c_2'). \tag{14}$$

So, substituting (14) in (11), we get the thesis:

$$\begin{aligned} \mathcal{L}(\mathcal{C}') - \mathcal{L}(\mathcal{C}) &\leq p(c_2)d(c_2, c_2') - p(c_1')d(c_1', c_1) + p(x)d(x, c_2') \\ &\quad - p(x)d(x, c_2') - p(c_2')d(c_2', c_2) - p(c_2)d(c_2, c_2') \\ &= -p(c_1')d(c_1', c_1) - p(c_2')d(c_2', c_2) \leq 0 \end{aligned} \tag{15}$$

□

Equation (15) assures that, if $x \in c_i$ and there exists some $j \neq i$ such that

$$d(x, c_j) < d(x, c_i), \tag{16}$$

then, moving the context x from class c_i to c_j will reduce the mutual information loss of the context classification. This property is at the basis of the MCECQ algorithm (proposed in [18, 19] and used in [20] and [24]) which consists in an iterative scanning of all contexts. For the current context x, its distance from its class $d(x, c_i)$ and from all other classes $d(x, c_j)$ are computed. If for one or more index the condition (16) is satisfied, the context x is moved to the "nearest" class. Proposition 3 assures that this algorithm always improves the mutual information loss (7). However, Eq. (11) shows clearly that it is not necessarily the best choice, as it does not take into account the true variation of $\mathcal{L}$, but only a quantity which has the same sign of it.

In conclusion, the result of proposition 2 shows that the popular algorithm based on MCECQ and used in [18, 20, 24], does not follow the *steepest descent* path towards the cost function minimum. Some explication is worth about the use of the expression "steepest descent" in this work. With this phrase we mean the single algorithm step which assures the maximal reduction of the cost function with respect to all the possible steps, *i.e.* with respect to all the possible context displacements from one class to another.

The proposition 2 suggests as well *how* the MCECQ algorithm should be modified in order to follow this optimal path: let $c_{f(x)}$ be the class which the context x belongs to according to the current classification function f, let c_k be a generic class, $c'_{f(x)} = c_{f(x)} - \{x\}$ and $c'_k = c_k \cup \{x\}$. Then we define the function δ as:

$$\delta(x,k) = p(c'_{f(x)})d(c'_{f(x)}, c_{f(x)}) - p(c_k)d(c_k, c'_k) + p(x)[d(x, c_{f(x)}) - d(x, c'_k)] \quad (17)$$

This function gives the mutual information variation associated to the displacement of the context x from its current class to the class c_k. The algorithm that we propose keeps moving the contexts according to the function δ until the relative variation of MI loss (indicated with ϑ) is smaller than a given threshold ϵ. More precisely, the algorithm has the following steps.

- Choose a starting classification function f, set $\mathcal{L}_{\text{old}} = \mathcal{L}(f)$ and set ϑ to any value greater than ϵ
- While $\vartheta > \epsilon$
  - for each $x \in \mathcal{X}$
    1. set $d_{\max}(x) = 0$
    2. for each $k \in \{1, 2, \cdots, F\}$
       - $\diamond$ if $\delta(x,k) > d_{\max}(x)$
         (a) set $d_{\max}(x) = \delta(x,k)$
         (b) set $f(x) = k$
       - $\diamond$ endif
    3. endfor $k \in \{1, 2, \cdots, F\}$
  - endfor $x \in \mathcal{X}$
  - Set $\vartheta = \frac{\mathcal{L}_{\text{old}} - \mathcal{L}(f)}{\mathcal{L}_{\text{old}}}$ and set $\mathcal{L}_{\text{old}} = \mathcal{L}(f)$
- Exit

We note that, since the algorithm only moves context from a class to another, if the initial classification is a partition of $\mathcal{X}$, then each new classification produced with this algorithm will be a partition as well. We call this algorithm MINIMA (Mutual INformation IMprovement Algorithm). We observe that within the inner loop over the class index k, we compare all the F partitions that we would obtain by moving x to each of the classes (by comparing the displacement to the current class with the best displacement

so far), and so at the end of loop the best move is kept. We would obtain the same result if at each k we would not move x from the original class and only at the end of the loop we would choose the best displacement.

On the other hand, at the end of the loop over the class index, the displacement of x is actually performed, and this influences the displacement of the next context (outer loop): we update the partition f and this new partition is used to decide about the displacement of the next context.

So we have an algorithm which is optimal with respect to the displacement of a given context, but at the same time is "greedy" since it moves the context x before evaluating the effect of this action over the displacement of the others. Therefore we have no assurance that MINIMA will produce the best CQ: the only way to find it would be to compute $\mathcal{L}$ for any possible partition, which is extremely demanding in terms of computation, and in practice, impossible. The only exception is the binary alphabet case, and some other special distribution that we will discuss in the next Section.

It is interesting to observe that MINIMA is brought back to the MCECQ if we use the following definition of δ.

$$\delta(x, k) = d(x, c_{f(x)}) - d(x, c_k) \tag{18}$$

We remark that both the strategies for δ assure a reduction of $\mathcal{L}(f)$ at each step, but only with the choice (17) we are assured to achieve the maximal reduction at each step. If we follow this trajectory we can hopefully find a better local minimum than what we can find with a non-optimal trajectory, but of course, we are not assured of it. Aware of this issue, we performed many experiments in order to validate MINIMA. Some of them (but not all, for the sake of brevity) are reported in Section 6, but here we underline that in every single experiment MINIMA always achieved better cost function values than classical algorithm based on (18). In other words, starting from the same initial conditions (*i.e.* the same initial classification function), MINIMA has always given lower cost-function final values than MCECQ.

5. Model-based classification algorithm

As well as its predecessors [18, 20, 24], the proposed algorithm has the problem that it can only find local minima of $\mathcal{L}(f)$. On the other hand, in [16], and then in [18] and [24], it is shown that if the input alphabet is binary, a dynamic programming (DP) algorithm can exactly solves the minimization problem, *i.e.* it is able to determine the global minimum of $\mathcal{L}(f)$.

To demonstrate this, we note that in the binary case the contexts can be ordered according to the conditional probabilities $\Pr(Y = 0|x)$ in a one-dimensional space. The crucial property is that in the optimal classification scheme, only consecutive contexts are merged in the same group. The demonstration of this property has been given in [19] and sketched in [24] We give here an alternative and complete formulation.

Let $p_i = \Pr(Y = 0|x_i)$. We have that:

$$p_1 < p_2 < p_3 \Rightarrow \begin{cases} d(x_1, x_2) < & d(x_1, x_3) \\ d(x_2, x_3) < & d(x_1, x_3) \end{cases} . \tag{19}$$

This is demonstrated observing that:

$$d(x_i, x_j) = d_2(p_i, p_j),$$

where $d_2(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ is the relative entropy between the binary distributions of parameters p and q. The first inequality in (19) is a consequence of the fact that

$$\frac{\partial d_2}{\partial q}(p, q) = \frac{p - q}{q(1 - q) \ln 2},$$

and so d_2 is increasing [decreasing] with q for all $q > p$ [$q < p$]. The second inequality in (19) can be demonstrated likewise. In conclusion, if $p_1 < p_2 < p_3$, then merging x_1 with x_2 or x_2 with x_3 is better that merging x_1 with x_3, that is, only contexts which are consecutive with respect to the p_i parameter, are merged together in the optimal classification. This reduces our problem to that of finding a set of thresholds in [0,1]. This is a one-dimensional problem, which can be resolved exactly with DP techniques.

This result is interesting because it allows to find the global minimum of the cost function, but it only holds for the binary case. In order to verify whether it can be extended to M-ary distributions, we tried to understand in which conditions one can use the DP technique. First, we observe that we need to sort the conditional pmf's with respect to some parameter λ_x (*e.g.* the probabilities $\Pr(Y = 0|x)$ in the binary case). We also note that a condition on contexts' merging, similar to (19), should hold. This condition should hold not only when merging single contexts, but also when classes are involved. At this end, a sufficient condition is that the conditional pmf's $p(y|c)$ have the same shape of $p(y|x)$. We remember that $p(y|c) = \sum_{x \in c} \frac{p(x)}{p(c)} p(y|x)$, so

we conclude that the conditional pmf's should have a *mixture invariance* property. In conclusion, we have found some *sufficient conditions* for a set of M-ary conditional probabilities so that the globally optimal classification function could be determined by a 1-D search. These are summarized in the following proposition.

Proposition 4. *If*

1. *the conditional pmf's depend from the context x only by a scalar parameter λ_x: $p(y|x) = q(y, \lambda_x)$;*
2. *for any class c there exists a value λ_c such that the conditional pmf can be expressed as: $p(y|c) = q(y, \lambda_c)$;*
3. *the pmf's $q(\cdot, \cdot)$ have relative entropies such that*

$$\lambda_1 < \lambda_2 < \lambda_3 \Rightarrow \begin{cases} D\left(\, q(y,\lambda_1) \;||\; q(y,\lambda_2) \,\right) < D\left(\, q(y,\lambda_1) \;||\; q(y,\lambda_3) \,\right) \\ D\left(\, q(y,\lambda_2) \;||\; q(y,\lambda_2) \,\right) < D\left(\, q(y,\lambda_1) \;||\; q(y,\lambda_3) \,\right) \end{cases} \tag{20}$$

then, the optimal classification can be found by a DP algorithm, as in the binary case.

Proof. When these hypotheses hold, all the conditional pmf's (for contexts and classes) are expressed by $q(\cdot, \cdot)$ functions. So we can sort the pmf's according to the parameter λ. Since the pmf q is mixture-invariant (*i.e.* condition 2 holds), even the merging of classes can be done according to the same principle. So we can adopt the same DP optimization algorithm used in the binary case. □

This generalization allows to find the globally optimal classification function even when the input alphabet is not binary. We just have to use the DP algorithm considered in the binary case, having λ instead of the probability $p(Y = 0)$ as optimization parameter. We give in the following an example where proposition 4 is applicable. Let us define, for each $y \in \{0, 1, \ldots, M-1\}$ and for each $\lambda \in (0, 1)$ the function:

$$\rho(y, \lambda) = \begin{cases} \lambda & \text{if } y = 0 \\ \frac{1-\lambda}{M-1} & \text{if } y \in \{1, \ldots, M\} \end{cases} \tag{21}$$

We use ρ to model the conditional probabilities, setting

$$p(y|x) = \rho(y, \lambda_x). \tag{22}$$

Therefore the pmf is parameterized with respect to the probability λ_x of the symbol 0. The other symbols are equiprobable. This is a quite particular model, but however it can apply in some interesting cases. In facts, it has arisen by observing the statistics of the motion segmentation maps described in the experimental section. We observe however that the proposed model is very similar to a binary one. When the input symbols have this kind of distribution, the following proposition holds.

Proposition 5. *Let the conditional probability mass functions $p(y|x)$ be in the form of $\rho(y, \lambda_x)$ $\forall x \in \mathcal{X}$.*

Then the globally optimal classification function can be found by the DP algorithm.

PROOF. Let $\lambda_c = E_{X \in c}[\lambda_X]$ and let $d_2(p, q)$ be the relative entropy between the binary distributions of parameters p and q. We first show that for any context x and for any class c,

$$p(y|c) = \rho(y, \lambda_c) \tag{23}$$
$$d(x, c) = d_2(\lambda_x, \lambda_c) \tag{24}$$

We know that $p(y|c) = \sum_{x \in c} \frac{p(x)}{p(c)} p(y|x)$. Using the definition of ρ, we have:

$$p(y|c) = \sum_{x \in c} \frac{p(x)}{p(c)} \lambda_x \delta(y) + \sum_{x \in c} \frac{p(x)}{p(c)} \sum_{k=1}^{M-1} \frac{1 - \lambda_x}{M-1} \delta(y - k) \tag{25}$$

We set:

$$\lambda_c = E_{X \in c}[\lambda_X] = \sum_{x \in c} \frac{p(x)}{p(c)} \lambda_x \tag{26}$$

We observe that:

$$\begin{aligned} \sum_{x \in c} \frac{p(x)}{p(c)} \frac{1 - \lambda_x}{M-1} &= \frac{1}{(M-1)p(c)} \left[p(c) - \sum_{x \in c} p(x) \lambda_x \right] \\ &= \frac{1}{M-1} \left[1 - \sum_{x \in c} \frac{p(x)}{p(c)} \lambda_x \right] \qquad = \frac{1 - \lambda_c}{M-1} \end{aligned} \tag{27}$$

Replacing (26) and (27) in (25), we get (23), which proves the mixture invariance property. On the other hand we get:

$$\begin{aligned} d(x,c) &= D\left(\,\rho(y,\lambda_x)\;||\;\rho(y,\lambda_c)\,\right) \\ &= \sum_{y=0}^{M-1} \rho(y,\lambda_x)\log\frac{\rho(y,\lambda_x)}{\rho(y,\lambda_c)} \\ &= \lambda_x \log\frac{\lambda_x}{\lambda_c} + (M-1)\frac{1-\lambda_x}{M-1}\log\frac{1-\lambda_x}{M-1}\frac{M-1}{1-\lambda_c} \\ &= \lambda_x \log\frac{\lambda_x}{\lambda_c} + (1-\lambda_x)\log\frac{1-\lambda_x}{1-\lambda_c} \end{aligned}$$

This assures that the relative entropies between conditional pmf's fulfill Eq. (20). It suffices to apply the same arguments that justify Eq. (19), since the relative entropy has the same structure.

Since all the hypotheses of proposition 4 hold, we can conclude with the thesis.□

The pmf described by (21) is not the only possible distribution satisfying the hypotheses of proposition 4, but the others we found share with it an essentially binary nature. However this model can result useful for non-binary signal that are sufficiently well described by one of these distribution. An example is given in the experimental section.

6. Experimental Results

6.1. Input data

Even though the MINIMA algorithm is general, we validate it for a particular problem, that is the encoding of the segmentation maps produced by a variation of the region-based (RB) video segmentation algorithm proposed in [12]. We want to compact these maps by a context-based arithmetic encoder, and in this case a context quantization step is needed, since the input alphabet is quite large.

Before giving the experimental results, it is useful to describe the main characteristics of the data to encode, *i.e.* the segmentation maps. The segmentation algorithm performs a block-based motion estimation (ME). The ME algorithm produces a segmentation map which tells for each macroblock (MB) whether it is composed by one or two regions, and in the latter case, it gives the initial and final points of the straight-line border between them.

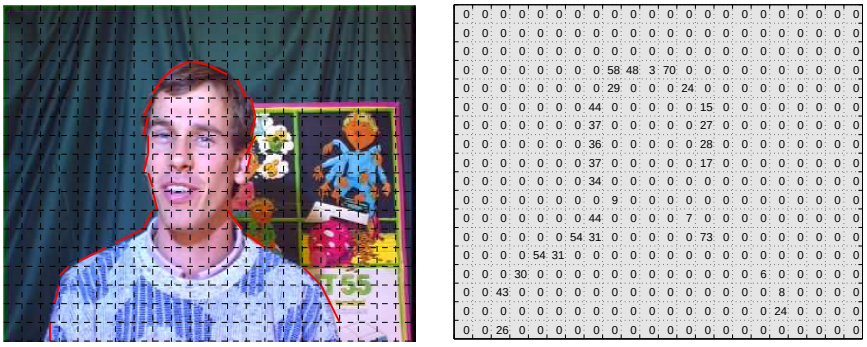

Figure 1: A sample segmentation map and its representation as bi-dimensional data set.

Preliminary studies prove that a finer description of these borders would have a high coding cost without a significant reduction of the motion-compensated error. Therefore we can constrain the initial and final point of each border segment to lie on only 5 positions per MB side (the corners, the middle point and two intermediate points). It is easy to show that in this case 80 different MB splits are possible, and so we represent each split with a numeric label in $\{1, 2, \ldots, 80\}$ and the one-region case with the label 0. The segmentation map is then a bi-dimensional array storing for each MB a label which tells if the MB is split or not, and, in the former case, which is the border between the two regions. The segmentation maps are then bi-dimensional sets of data, with a large input alphabet $\mathcal{Y} = \{0, 1, \ldots, 80\}$, and with strong but highly non-linear dependencies among samples. An example of segmentation map produced by this algorithm, and the associated label array are in Fig. 1. We note that in this case, context-based entropy coding will only work with context quantization, since the alphabet is relatively large, and the statistical dependence among data is strong, but not linear.

6.2. Coding results

We used the RB ME algorithm to produce the segmentation maps for several popular test sequences, as "eric", "flower and garden", "foreman", "mother and daughter", "paris" and "silent". These sequences differ one another in motion content, as some of them are quite static ("head and shoulders" sequences), others present regular motion, others have irregular motion content. The segmentation maps were to be encoded by a context-based arithmetic encoder, with the causal template shown in Fig. 2. With

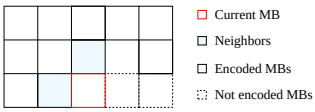

Figure 2: The causal template used for the experiments.

such a neighborhood, the number of possible different contexts would be as high as $81^2 = 6561$, so context quantization was necessary.

In a first set of experiments, we used two context quantization algorithms to reduce the number of conditioning states to 20: MCECQ, proposed in [19], (or the equivalent algorithms in [20] and in [24]), and MINIMA. As highlighted in section 4, they share the same structure, and only differ in the definition of the δ function, which is given by (17) for our algorithm and by (18) for the other one.

The algorithms start from the same initial classification function, chosen randomly, and then the iterative optimization is performed. The probability functions needed to run the algorithms were estimated on a training set obtained from the test sequences[5]. In order to reduce the dependence on the starting configuration, the entire process (random initialization and iterative optimization) was repeated 20 times and the best resulting classification function was retained for both algorithms. We note that in each iteration, MINIMA has reached a lower cost function value than MCECQ. This was already an interesting experimental result, since just having optimized the iterative step within the research algorithm does not assure that a better local minimum will be found. However, we underline that we carried out similar experiments for several other sequences, with many different parameter configurations, and, in any test we performed, MINIMA attained a lower cost function than MCECQ. These results are encouraging about the efficiency of the proposed algorithm in a general case.

In order to complete the comparison, we ran a context-based arithmetic encoder using the context classifications obtained with both the algorithms. In both cases, the CQ was embedded into the encoder, and they are known at the decoder side as well. However, even if the CQ should be sent to the decoder, its cost is fixed and independent from the data length, so it becomes

[5]The data used as training set were not used again to evaluate the compression performance

Sequence	Coding rate [bits/sample] MINIMA	MCECQ
"eric"	1.61	1.67
"flower and garden"	3.22	3.25
"foreman"	3.32	3.33
"mother and daughter"	0.94	0.95
"paris"	1.50	1.52
"silent"	1.38	1.39

Table 1: Coding rates [bits/sample] achieved with the proposed and the MCECQ [18] algorithms.

negligible if enough data are coded.

The achieved coding rates are shown in Tab. 1. These results tell that the proposed algorithm allows some gain, even though the difference with MCECQ is often small. Afterwards an extensive experimentation, which also comprised stochastic relaxation methods, we believe that the small difference between the two algorithms is ascribable to the fact that in this case the mutual information loss $\mathcal{L}(f)$ has many local minima, but all of them at quite close levels.

As far as the choice of the optimal number of classes is concerned, we repeated the above experiment for each F from 2 to 200. The MINIMA algorithm is consistently better than MCECQ (the obtained results are similar to those of Tab. 1), and we noticed that little or no gain is obtained when more than 50 conditioning contexts are used, and that anyway performances do not change a lot for $F > 20$. We conclude that for a given kind of data to be coded, a good value for F can be determined by off-line preliminary tests, and then the MINIMA algorithm can be used to find the corresponding optimal classification.

6.3. The model-based algorithm

In a second set of experiments, we tested the model-based algorithm described in section 5, which consists in applying a DP algorithm to the scalar parameters of a set of distributions like the one in (22). In facts, the DP model-based algorithm was suggested just by the statistical analysis of segmentation maps. It results that, for many contexts x, as for example the one illustrated in Fig. 3(a), the symbol "0" (meaning no split, *i.e.* one region in the MB) is quite more probable than all other symbols, which in their turn result almost equiprobable. Therefore, the conditional probability

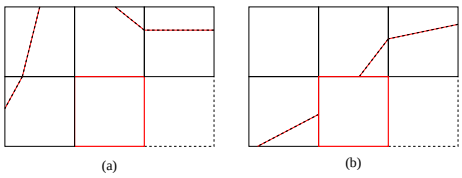

Figure 3: Segmentation maps: examples of contexts for the current MB (in red).

of the segmentation symbols $p(Y|x)$ has a structure like in (22). However, we are aware that this model does not perfectly match our signal: if we consider a context as the one in Fig. 3(b), it is reasonable that some non-zero symbols are more probable than others (namely those representing North-West border). We underline that the segmentation maps have just inspired our effort towards a model-based algorithm which could attain the global optimum even for non-binary sources, but they are not completely fit with the statistics described by Eq. (22). As a consequence we do not expect a large improvement of performances using the model-based algorithm on the segmentation maps.[6]

On the other hand, we validated the model-based algorithm with synthetic data generated so that their statistics exactly match the proposed model. This is of course a favorable case for the model-based algorithm, but anyway it provides some useful insights about the potentialities of the method. In particular, we used a bi-dimensional random process with conditional distributions described by equation (22). The input alphabet has a variable cardinality (values as $M = 4, 6, 8, 10, 15$ have been tested), and the template is made up by the left and the upper sample (as shown in Fig. 2). We used MINIMA and the model-based algorithm to look for the minimal point of the cost function $\mathcal{L}$, with a number of classes F variable from 2 to $M^N - 1$. In this case the model-based algorithm is able to find the global optimum, and so the corresponding value $\mathcal{L}^*$ of the cost function is lower than the local minimum $\mathcal{L}_0$ we can find with MINIMA. In particular we computed the relative reduction $\eta = \frac{\mathcal{L}_0 - \mathcal{L}^*}{\mathcal{L}_0}$. The results are shown in Fig. 4, where on the horizontal axis there is the ratio between the number of class F and the number of contexts M^N, so that we can report on the same graph different values of M. On the vertical axis there is the relative reduction η. We note that η consistently increases with the number of classes, irrespectively to the

[6]This is confirmed by some preliminary tests we have performed.

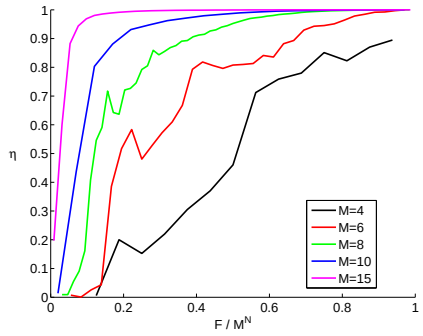

Figure 4: Cost function relative improvement of the DP model-based algorithm with respect to MINIMA for the synthetic data.

alphabet cardinality M. This can be interpreted as an evidence that the optimal classification becomes harder when the number of contexts increases, and so an algorithm which can only find the local minimum is disadvantaged. For larger M's, $\eta(F/M^N)$ grows faster because for the same ratio F/M^N the actual number of classes is higher.

However it should be noted that while the relative difference between $\mathcal{L}^*$ and $\mathcal{L}_0$ increases with F, as expected, their absolute difference (together with their values) decreases and tends to zero for F tending to M^N. As an example, we report in Tab. 2 some experimental values of $\mathcal{L}^*$ and $\mathcal{L}_0$ for $M = 15$ and for various values of F.

As a final evaluation of the model-based algorithm, we computed the impact of the different classifications over the arithmetic encoder rate. Of course, the coding rate gain is less than η, since theoretically $R = H(Y|X) + \mathcal{L}$. Yet, we found a non-negligible rate reduction over our synthetic data sequences, as shown in Tab. 3. These results allow us to conclude that the model-based approach is worth when the input signal statistics are exactly expressed by (21), while when this equation holds only approximately, it is difficult to foretell which algorithm between MINIMA and the DP model-based one will have the better performance.

7. Conclusions

In this work we review the current methods for solving the important problem of context quantization, which arises when data sets with large alphabet or large causal template are to be encoded with context-based entropy

Classes	$\mathcal{L}_0$	$\mathcal{L}^*$
2	0.097	0.080
5	0.070	0.010
7	0.063	0.005
10	0.057	0.003
15	0.047	0.002
20	0.040	0.001
25	0.034	< 0.001
30	0.029	< 0.001
35	0.024	< 0.001
40	0.021	< 0.001
50	0.015	< 0.001
60	0.011	< 0.001

Table 2: Cost function values [bits] achieved by MINIMA and by the DP model-based algorithm, for $M = 15$.

M	MINIMA	DP Model-Based	Gain
4	1.85	1.76	4.9%
6	2.15	2.13	0.9%
8	2.26	2.23	1.3%
10	2.51	2.45	2.4%
15	3.20	3.10	3.1%

Table 3: Coding rates [bit per sample] for synthetic data.

methods. We provide a detailed mathematical analysis of the problem, which sheds light on some fine details of it. This allows a fine comprehension of the process of context grouping, and also to discover a suboptimal step in the popular MCECQ algorithm proposed in [19], [20], and [24]. Moreover our analysis directly suggests how to modify this algorithm to follow the steepest descent path in the minimization of the cost function. Additionally, we show that the global minimum is achievable not only for binary-distributed input symbols, but also for M-ary distributions, provided that some conditions are fulfilled. This happens for example when the conditional probabilities can be expressed as in (21). This model has been suggested by the motion segmentation map problem.

Even though this work is rather a theoretical one, we perform some experiments in order to look for confirmation to the proposed techniques. We use our algorithm in order to find a context classification for the encoding of

some video segmentation maps produced by a region-based ME algorithm. The peculiarity of these data prevents an efficient use of generic image lossless coding algorithm, because of the complex non-linear dependencies. In this case, MINIMA provides slightly better results than MCECQ. We also try to apply the model-based version, and encouraging results are obtained in the case of a synthetic signal produced accordingly to our statistical model.

Future work will be devoted to use the proposed algorithm into a complete video coder using segmentation maps in order to perform an accurate motion compensation. In this context, an efficient coding of the segmentation map is of crucial importance in order to achieve competitive rate-distortion performances. We are also investigating several models for conditional probabilities, in order to determine whether there exist other cases in which the DP approach can be used to find the global minimum of the cost function $\mathcal{L}$.

References

[1] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1309–1324, Aug. 2000.

[2] *Digital Compression and Coding of Continuous-Tone Still Images*, ISO/IEC JTC1, ITU-T, Sep. 1992, ISO/IEC 10918-1 — ITU-T Recommendation T.81.

[3] *Information Technology — JPEG 2000 Image Coding System*, International Standardization Organization, Jul. 2002, ISO/IEC 15444.

[4] *Coding of moving pictures and associated audio for digital storage mMedia at up to about 1.5 Mbps*, ISO/IEC JTC1, 1993, ISO/IEC 11172-2.

[5] *Generic coding of moving pictures*, ISO/IEC JTC1, 1995, ISO/IEC 13818-2.

[6] *Coding of audio-visual objects*, ISO/IEC JTC1, Apr. 2001, ISO/IEC 14496-2.

[7] *Advanced video coding for generic audiovisual services*, International Telecommunication Union - Telecommunication Standardization Sector, Mar. 2005, ITU-T Recommendation H.264.

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

[9] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression*. Upper Saddle River, NJ: Prentice-Hall, 1990.

[10] X. Wu and N. D. Memon, "Context-based, adaptive, lossless image coding," *IEEE Transactions on Communications*, pp. 437–444, Apr. 1997.

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

[12] S. Boltz, É. Debreuve, and M. Barlaud, "A joint motion computation and segmentation algorithm for video coding," in *Proceedings of European Signal Processing Conference*, Antalya, Turkey, Sep. 2005.

[13] M. J. Weinberger, J. J. Rissanen, and R. B. Arps, "Applications of universal context modeling to lossless compression of gray-scale images," *IEEE Transactions on Image Processing*, vol. 5, no. 4, pp. 575–586, 1996.

[14] X. Wu, "Lossless compression of continuous-tone images via context selection, quantization and modeling," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 656–664, May 1997.

[15] C. Chrysafis and A. Ortega, "Efficient context-based entropy coding for lossy wavelet image compression," in *Proceedings of Data Compression Conference*, Snowbird, USA, Mar. 1997.

[16] D. Green, F. Yao, and T. Zhang, "A linear algorithm for optimal context clustering with application to bi-level image coding," in *Proceedings of IEEE Internantional Conference on Image Processing*, vol. 1, Oct. 1998, pp. 508–511.

[17] A. Bookstein, S. T. Klein, and T. Raita, "An overhead reduction technique for mega-state compression schemes," in *Proceedings of Data Compression Conference*, Snowbird, USA, Mar. 1997, pp. 367–376.

[18] X. Wu, P. A. Chou, and X. Xue, "Minimum conditional entropy context quantization," in *Proceedings of IEEE International Symposium on Information Theory*, Sorrento, Italy, Jun. 2000, p. 43.

[19] ——, "Minimum conditional entropy context quantization," 2000. [Online]. Available: http://research.microsoft.com/ ~pachou/pubs/WuCX00L.ps.gz

[20] J. Chen, "Context modeling based on context quantization with application in wavelet image coding," *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 26–32, Jan. 2004.

[21] S. Forchhammer, X. Wu, and J. D. Andersen, "Optimal context quantization in lossless compression of image data sequences," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 509–517, Apr. 2004.

[22] ——, "Lossless image data sequence compression using optimal context quantization," in *Proceedings of Data Compression Conference*, Snowbird, USA, Mar. 2001, pp. 53–62.

[23] S. Forchhammer and X. Wu, "Context quantization by minimum adaptive code length," in *Proceedings of IEEE International Symposium on Information Theory*, Nice (France), Jun. 2007, pp. 246–250.

[24] Z. Liu and L. J. Karam, "Mutual information-based analysis of JPEG2000 contexts," *IEEE Transactions on Image Processing*, vol. 14, no. 4, pp. 411–421, Apr. 2005.

[25] X. Delaunay, M. Chabert, G. Morin, and V. Charvillat, "Bit-plan analysis and contexts combining of JPEG2000 contexts for on-board satellite image compression," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HA, Apr. 2007.

[26] A. Katsaggelos, L. P. Kondi, F. W. Meier, J. Ostermann, and G. M. Schuster, "MPEG-4 and rate-distortion-based shape-coding techniques," *Proceedings of the IEEE*, no. 6, pp. 1126–1154, Jun. 1998.

[27] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, Jul. 2003.