A GLOBAL APPROACH TO JOINT QUANTIZER DESIGN FOR DISTRIBUTED CODING OF CORRELATED SOURCES

Ankur Saxena[*]*, Jayanth Nayak*[†] *and Kenneth Rose*[*]

[*] ECE Department, University of California, Santa Barbara, CA 93106, USA.
[†]IRISA/INRIA, Campus Universitaire de Beaulieu, 35042, Rennes France.
Email:{ankur, rose}@ece.ucsb.edu, jnayak@irisa.fr

ABSTRACT

The focus of this work is on the design of efficient quantizers for correlated sources subject to complexity limitations on the encoding terminals. Existing iterative descent methods rely heavily on initialization, and the prevalence of numerous 'poor' local optima strongly motivates the use of a global design algorithm. We propose a multiple-prototype based deterministic annealing approach for the joint design of all components of a generic distributed source coding system. Our approach avoids many poor local optima, is independent of initialization and does not assume any prior information on the underlying source distribution. Simulation results show substantial gains over a Lloyd-like iterative algorithm.

1. INTRODUCTION

Consider a distributed network of limited complexity sensors that transmit information to a central unit. Nearby sensors may be designed to observe different physical quantities, e.g., temperature, humidity, pressure. We are interested in efficiently reconstructing one or more physical quantities at the central unit (decoder). Typically, there is a high degree of correlation between data being transmitted by different sensors. Since the encoders at each sensor location function independently, they will not, in practice, accomplish the maximum possible lossy compression of the source pair. However, to obtain the best possible transmission rates from independent encoders, it is necessary that the code *design* is performed jointly.

The field of distributed source coding began in the seventies with the seminal work of Slepian & Wolf [9] who gave bounds for lossless coding of correlated sources. Later, Wyner & Ziv [11] extended the analysis to lossy coding. However, it was not until the late nineties that practical methods based on nested lattices [12] or channel coding principles [5] for designing quantizers were developed. From the source coding perspective, generalizations of the Lloyd algorithm [4] were presented in [2, 3, 7] where different encoders and decoders were optimized in an iterative fashion to design locally optimal quantizers.

Current approaches based on channel coding are generally suitable when sources can be modeled as noisy versions of each other, where the noise is additive and unimodal in nature. But such approaches are of limited use where such simplifying assumptions do not apply. An illustrative example is when, say, the temperature and humidity are drawn from a mixture of joint gaussian densities, where the mixture components are due to varying underlying conditions such as the time of day, pressure, etc. On the other hand, Lloyd-algorithm based methods to design a distributed vector quantizer (DVQ) suffer from the presence of numerous 'poor' local minima on the distortion-cost surface. Clever initializations such as the ones proposed in [10] for multiple description scalar quantizer design may lead to better or even global minima, but to the authors' knowledge there have been no generalization to vector quantization, nor has such a scheme been found for DVQ design. All these difficulties underline the need for a global optimization scheme, i.e., a powerful optimization tool that provides the ability to avoid poor local optima. We propose a deterministic annealing approach for optimal DVQ design.

Deterministic annealing (DA) is motivated by the process of annealing in physics. It is independent of the initialization, does not assume any knowledge about the underlying source distribution and avoids many poor local minima of the distortion-cost surface [8]. In DA, the encoding rule is randomized and the expected distortion is minimized subject to a constraint on the level of randomness as measured by the Shannon entropy of the system. The Lagrangian functional can be considered as the free energy of the system and the Lagrangian parameter as the 'temperature'. The minimization is started at high temperature (high degree of randomness), where, in fact the entropy is maximized and hence all the reproduction points are at the centroid of the source distribution. The minimum is then tracked at successively lower levels of entropy (temperature), by recalculating the optimum locations of the reproduction points and the encoding probabilities at each stage. As the temperature reaches zero, the average distortion cost is directly minimized and a deterministic encoder is obtained.

DA can also be used in the estimation of a signal from one or more of its noisy versions, e.g., efficient quantizers can be designed for the CEO problem [1]. A DA-based approach for estimating a signal from its corrupted version has been shown to offer significant improvement over other traditional methods [6].

The rest of the paper is organized as follows. In Section 2, we state the problem formally, establish the notation and describe an iterative method based on Lloyd's algorithm for multiple prototype (MP) coder design. In Section 3, we derive the DA approach to DVQ design and provide its update formulae (necessary optimality conditions). Experimental results are given in Section 4, followed by the Conclusions section.

[†]This work was carried out while J. Nayak was with the University of California, Santa Barbara.

The work is supported in part by the NSF under grant IIS-0329267, the University of California MICRO program, Applied Signal Technology, Inc., Dolby Laboratories Inc., and Qualcomm Inc.

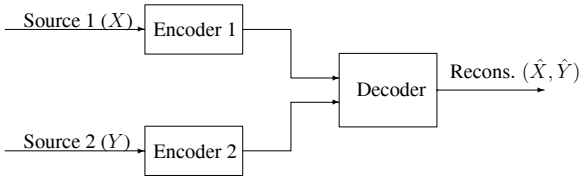

Fig. 1. Correlated Source Coding

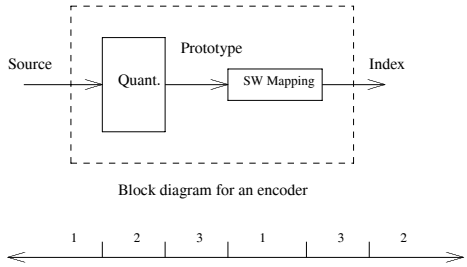

Fig. 2. Block diagram of an encoder and an example of SW mapping from prototypes (Voronoi regions) to indices.

2. PROBLEM STATEMENT AND ITERATIVE DESCENT METHODS

Consider the scenario in Fig. 1. X and Y are two continuous-valued i.i.d., correlated (possibly vector) sources which are compressed and transmitted independently at rates R_1 and R_2 bits per sample respectively by the encoders. The decoder wishes to reconstruct either one or both sources and minimize the following expected distortion:

$$E\{\alpha d(X, \hat{X}) + (1-\alpha) d(Y, \hat{Y})\}, \tag{1}$$

where $\alpha \in [0, 1]$, and $\hat{X}$ and $\hat{Y}$ are the reconstruction values for X and Y respectively. For this, we need to design encoders for X and Y, and a joint decoder. Design techniques based on iterative descent methods which converge to a local minimum have been proposed in the literature [2, 3, 7]. We next adopt this framework and describe a locally optimum algorithm for the multiple prototypes structure, which can be viewed as combining histogram or kernel based techniques for estimating source distributions and quantizer design. This approach will underline the need for powerful optimization schemes such as DA.

Specifically, we have a training set $\mathcal{T} \equiv \{\mathcal{X}, \mathcal{Y}\}$, which consists of m-dimensional i.i.d. vectors. We design a vector quantizer 'Q' for X using Lloyd's algorithm [4]. Q assigns training set data points to one of the $\mathcal{K}$ regions, C_k^x. The regions C_k^x partition the space into disjoint Voronoi regions each associated with a prototype x_k. Next, each one of the $\mathcal{K}$ regions C_k^x is mapped to one of the $\mathcal{I}$ indices, via a mapping $v(k) = i$, which we refer to as Slepian-Wolf (SW) mapping, since this mapping is the module that in fact exploits the correlation between sources. The index 'i' is finally transmitted to the central unit. The block diagram of the two stages of encoder for a source and an example of SW mapping with m =1, $\mathcal{K}$ = 6 and $\mathcal{I}$ = 3 is given in Fig. 2. The region associated with an index i is $R_i^x = \bigcup_{k;v(k)=i} C_k^x$.

We next define regions C_l^y, R_j^y and prototypes y_l in the Y domain, following the same steps that led to C_k^x, R_i^x and x_k. Here, the $\mathcal{L}$ regions are mapped to $\mathcal{J}$ indices via SW mapping $w(l) = j$. We re-emphasize that the idea behind the SW mappings is to exploit the correlation between the quantized versions of the sources and to reduce the transmission rate. Finally, at the decoder we have $\mathcal{I} \times \mathcal{J}$ reconstruction values $\hat{x}_{ij}$ and $\hat{y}_{ij}$ for X and Y respectively. To minimize the expected distortion defined in (1), the SW mappings v, w and reconstruction values $\hat{x}_{ij}$ and $\hat{y}_{ij}$, which are initialized 'randomly' are optimized iteratively using the following steps:

1. **SW Mapping for X**: For $k = 1 : \mathcal{K}$, assign region k to index i, i.e., $v(k) = i$ such that:

$$i = \arg\min_{i'} \sum_{\substack{(x,y)\in\mathcal{T};\\ x\in C_k^x}} \{\alpha d(x, \hat{x}_{i'j(y)}) + (1-\alpha) d(y, \hat{y}_{i'j(y)})\}. \tag{2}$$

2. **SW Mapping for Y**: For $l = 1 : \mathcal{L}$, assign region l to index j, i.e., $w(l) = j$ such that:

$$j = \arg\min_{j'} \sum_{\substack{(x,y)\in\mathcal{T};\\ y\in C_l^y}} \{\alpha d(x, \hat{x}_{i(x)j'}) + (1-\alpha) d(y, \hat{y}_{i(x)j'})\}. \tag{3}$$

3. **Decoder Rule**: For $i = 1 : \mathcal{I}$ and $j = 1 : \mathcal{J}$, find $\hat{x}_{ij}$ and $\hat{y}_{ij}$, such that:

$$\hat{x}_{ij} = \arg\min_{a} \sum_{x\in R_i^x, y\in R_j^y} d(x, a), \tag{4}$$

$$\text{and} \quad \hat{y}_{ij} = \arg\min_{b} \sum_{x\in R_i^x, y\in R_j^y} d(y, b). \tag{5}$$

We will refer to this approach as Separate-Lloyd (SL) because initial quantizers are separately designed using Lloyd's algorithm and fixed. Then the mappings from prototypes to indices for X and Y and the final reconstruction values are optimized in an iterative manner. The SL approach inherits from Lloyd's algorithm the interrelated shortcomings of getting stuck in a local minima, and dependence on initialization. Moreover, the source correlation has been ignored during the design of quantizers for the individual sources. All these issues call for the use of a global optimization scheme, such as DA. We present the DA algorithm, its necessary conditions for optimality and finally the simulation results.

3. DERIVATION OF THE DA ALGORITHM

3.1. Preliminaries

A formal derivation of the DA algorithm is based on principles borrowed from information theory and statistical physics. Here the deterministic encoder is replaced by a random encoding rule. For a detailed derivation of DA, please refer [8].

Let us first consider the initial quantizer for X. The original training set data point is assigned to each prototype in probability. These probabilities are determined by finding the distribution that minimizes an appropriately defined distortion cost between data point and prototype, subject to a specified level of randomness (measured by Shannon conditional entropy of the encoding probability distribution). Alternatively, we can view the prototypes as partitioning the space into Voronoi regions, so a structural constraint is imposed on the association probabilities. The structural cost function that imposes the desired partition is:

$$D_1 = \frac{1}{N} \sum_k \sum_{(x,y)\in\mathcal{T}; x\in C_k^x} d(x, x_k), \tag{6}$$

whose probabilistic equivalent is the expected structural cost:

$$\langle D_1 \rangle = \frac{1}{N} \sum_k \sum_{(x,y)\in\mathcal{T}} c_{k|x} d(x, x_k), \tag{7}$$

where $c_{k|x} = \Pr[x_k|x] = \Pr[x \in C_k^x]$ is the probability of quantizing data point x to prototype x_k and hence to the k^{th} 'Voronoi' region. N is the number of points in the training set.

We choose the distribution that minimizes $\langle D_1 \rangle$ subject to a constraint on the Shannon entropy,

$$H_1 = \frac{-1}{N} \sum_k \sum_{(x,y)\in\mathcal{T}} c_{k|x} \log(c_{k|x}). \tag{8}$$

Minimizing $\langle D_1 \rangle$ subject to a constraint on H_1 yields the Gibbs distribution,

$$c_{k|x} = \frac{e^{-\gamma_1 d(x,x_k)}}{\sum_m e^{-\gamma_1 d(x,x_m)}}, \tag{9}$$

where γ_1 is the inverse 'temperature' controlling the 'fuzziness' in the X quantizer.

Similarly in the Y domain, we have $c_{l|y} = \Pr[y_l|y] = \Pr[y \in C_l^y]$. Here we minimize $\langle D_2 \rangle$ subject to the constraint, H_2 to get the encoding probabilities $c_{l|y}$. The respective expressions are:

$$\langle D_2 \rangle = \frac{1}{N} \sum_l \sum_{(x,y)\in\mathcal{T}} c_{l|y} d(y, y_l), \tag{10}$$

$$H_2 = \frac{-1}{N} \sum_l \sum_{(x,y)\in\mathcal{T}} c_{l|y} \log(c_{l|y}), \tag{11}$$

$$c_{l|y} = \frac{e^{-\gamma_2 d(y,y_l)}}{\sum_p e^{-\gamma_2 d(y,y_p)}}. \tag{12}$$

Again, here γ_2 controls the fuzziness in the Y quantizer.

We now recall that $v(k)$ and $w(l)$ are the SW mappings from the prototypes to the indices. The Gibbs distribution over Voronoi cells induces the following distribution of encoding to the possibly non-contiguous 'regions' associated with the indices 'i' and 'j':

$$r_{i|x} = \sum_{k:v(k)=i} c_{k|x}, \tag{13}$$

$$r_{j|y} = \sum_{l:w(l)=j} c_{l|y}. \tag{14}$$

Note that so far we have only considered the structural distortion that imposes a Voronoi structure on the prototype partition. The overall distortion function which we seek to minimize is:

$$D = \frac{1}{N} \sum_{k,l} \sum_{(x,y)\in\mathcal{T}} c_{k|x} c_{l|y} \{\alpha d(x, \hat{x}_{ij}) + (1-\alpha) d(y, \hat{y}_{ij})\}, \tag{15}$$

subject to a constraint on the *joint entropy* of the system.

By construction the source variables X and Y and the transmitted indices I and J form the Markov chain: $I - X - Y - J$. Hence the joint entropy of the system is $H(X, Y, I, J) = H(X, Y) + H(I|X) + H(J|Y)$, where $H(X, Y)$ is the source entropy, which is obviously unchanged by encoding decisions. Note that these entropies are different from those imposed on structural costs in (7) and (10) to obtain Gibbs distribution in (9) and (12) respectively, since the latter's objective was imposing MP based nearest-neighbor structure on the initial quantizers. The optimization of D in (15) subject to the entropy constraint is equivalent to the Lagrangian minimization,

$$\min_{\{x_k\},\{y_l\},\gamma_1,\gamma_2,\{\hat{x}_{ij}\},\{\hat{y}_{ij}\}} L = D - TH \tag{16}$$

where the Lagrange parameter T (temperature) controls the entropy of the distribution.

3.2. Update Equations for DVQ Design

Here we specialize to the squared-error distortion measure to provide specific update formulae. The approach is clearly not restricted to this choice. At a fixed temperature T, the free energy L may be minimized by a *gradient descent* procedure using the following expressions for the gradients:

$$\frac{\partial L}{\partial \hat{x}_{ij}} = \frac{-2\alpha}{N} \sum_{(x,y)\in\mathcal{T}} r_{i|x} r_{j|y} (x - \hat{x}_{ij}), \tag{17}$$

$$\frac{\partial L}{\partial \gamma_1} = \frac{1}{N} \sum_{\substack{i,j,k: \\ v(k)=i}} \sum_{(x,y)\in\mathcal{T}} r_{j|y} c_{k|x} (x - x_k)^2 (L_{xy} - L_{xy}^{ij}) \tag{18}$$

$$\frac{\partial L}{\partial x_k} = \frac{2\gamma_1}{N} (x - x_k) c_{k|x} \sum_j r_{j|y} (L_{xy}^{v(k)j} - L_{xy}), \tag{19}$$

where $L_{xy}^{ij} = \alpha(x - \hat{x}_{ij})^2 + (1-\alpha)(y - \hat{y}_{ij})^2 + T \log \sum_{k:v(k)=i} e^{-\gamma_1 (x-x_k)^2} + T \log \sum_{l:w(l)=j} e^{-\gamma_2 (y-y_l)^2}$ is the contribution to the cost incurred when the data pair (x, y) is reconstructed by using indices i and j, received from the encoders of X and Y respectively. Also $L_{xy} = \sum_{i,j} r_{i|x} r_{j|y} L_{xy}^{ij}$ is the average contribution to the cost due to the data pair (x, y). The gradients with respect to $\hat{y}_{ij}$, γ_2 and y_l can be directly stated using symmetry. The annealing is started from a high temperature and is performed at a sequence of temperatures that are successively lower. At the limit of zero temperature, quenching is done, i.e., the scaling parameters γ_1 and γ_2 are forced to infinity and hard multiple prototype DVQ cost (1) is optimized.

4. EXPERIMENTAL RESULTS

We first illustrate the DA advantage in DVQ design via a toy example. Suppose there are two sensors which can detect a signal above a threshold and can transmit only one bit of information. To decide this threshold is equivalent to a design of two level quantizer. Let the signals at the sensors be $X \sim N(0, 1)$ and $Y = X + Z$ where $Z \sim N(0, 0.1)$ is independent of X and we are interested in the reconstruction of X only, i.e., α = 1. The rates are $R_1 = R_2 = 1$ bit and allow 1 prototypes per source X and Y. Design a DVQ using a 2000 point training set data. Using the SL approach, we get partitions for X and Y about their means (origin). Thus, the correlation between the sources cannot be exploited. The DA approach, on the other hand gives the thresholds as -0.47 and 0.78 for X and Y, respectively. The expected distortion that SL and DA yield is approximately -5 dB versus -6.7 dB in this example for the above mentioned partitions. This clearly shows the gains that DA can achieve over the SL approach.

In the next two examples, X and Y are drawn from a mixture of four joint gaussians. In our simulations, the mixtures components are assumed to be equiprobable. The means for X, Y and correlation coefficients for the four components are taken as {0, 0, 0.87}, {1, 0.5, 0.9}, {-1, 1, -0.92} and {2, -1, -0.95} respectively. The variance of X and Y in all the components of the mixture was taken to be 1.

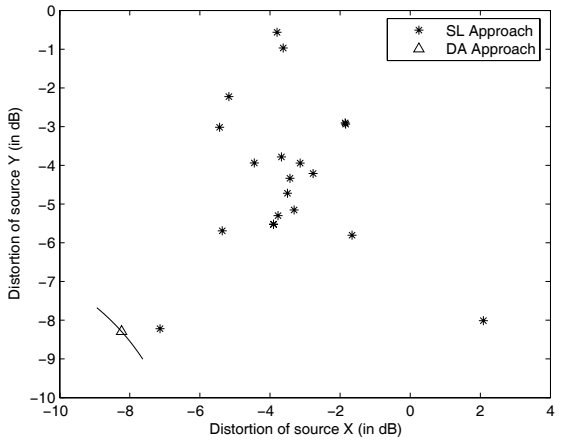

Fig. 3. Comparison between SL and DA approaches for R_1 = R_2 = 2, $\mathcal{K}$ = $\mathcal{L}$ = 4, α = 0.5. **D** from DA is -8.26 dB while SL gives best and worst **D** as -7.64 and -0.52 dB respectively. For ease of comparison, a line along which **D** = -8.26 dB is drawn.

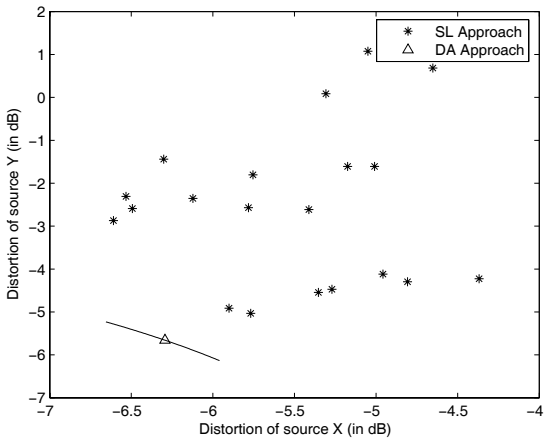

Fig. 4. Comparison between SL and DA approaches for R_1 = 2, R_2 = 1, $\mathcal{K}$ = 8, $\mathcal{L}$ = 4, α = 0.6. **D** from DA is -6.02 dB while SL gives best and worst **D** as -5.47 and -1.54 dB respectively. For ease of comparison, a line along which **D** = -6.02 dB is drawn.

The SL algorithm is run 20 times while joint-DA only once. In both examples, a scalar quantizer is designed using a training set of 2000 samples. In the first case, the weight factor α is taken to be 0.5 and the rates for X and Y are 2 bits each. A low complexity system is designed with 4 prototypes for both X and Y. The results are shown in Fig. 3. Here DA outperforms the best solution obtained by SL algorithm by $\sim$ 0.6 dB in terms of expected distortion. Note that there is a huge gap of $\sim$ 7.1 dB between the best and worst runs of SL algorithm. This can be explained by non-efficient SW mapping from the prototypes to indices in the second stage of SL technique. It highlights the fact that the cost surface is riddled with local minima that easily trap greedy techniques such as SL. In the second case, α is 0.6, R_1 = 2, R_2 = 1 while the number of prototypes for X and Y are 8 and 4 respectively. The results are shown in Fig. 4. Note that sometimes the distortion for X obtained from SL algorithm is better than DA but the net distortion achieved by DA is consistently better.

One might argue that if the initial quantizers are of high resolution and the number of prototypes/index is large, then we can compensate for the loss incurred in separate design of initial first level quantizers. However, this will merely transfer the difficulty into the SW mapping modules and exacerbate their design complexity, the results of which is critical as shown in the above example. Ongoing research is concerned with 'softening' the SW mappings as well which will further enhance the algorithm performance.

5. CONCLUSIONS

We have proposed a multiple prototype based deterministic annealing approach for the design of low resolution quantizers for correlated sources. This approach assumes no prior knowledge about the underlying probability distribution of the sources, eliminates the dependence on good initial configurations and avoids many poor local minima of the distortion cost surface. The necessary conditions (and update equations) for quantizer design are derived and presented. Experimental results comparing DA with Separate-Lloyd algorithm are shown. Significant improvements confirm the advantage of using a global optimization scheme such as DA for correlated sources quantizer design.

6. REFERENCES

[1] T. Berger, Z. Zhang and H. Viswanathan, "The CEO problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 887-902, May 1996.

[2] J. Cardinal and G. Van Assche, "Joint entropy-constrained multiterminal quantization," *Proc. IEEE ISIT,* p. 63, June 2002.

[3] M. Fleming, Q. Zhao and M. Effros, "Network Vector Quantization," *IEEE Trans. on Information Theory,* vol. 50, no. 8, pp. 1584-1604, Aug. 2004.

[4] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. on Information Theory,* vol. 28, pp. 129-137, March 1982.

[5] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *Proc. IEEE Data Compression Conference,* pp. 158-167, March 1999.

[6] A. Rao, D. Miller, K. Rose, A. Gersho, "A Generalized VQ method for combined compression and estimation," *Proc. ICASSP,* vol. 4, pp. 2032-2035, May 1996.

[7] D. Rebollo-Monedero, R. Zhang and B. Girod, "Design of optimal quantizers for distributed source coding," *Proc. IEEE Data Compression Conference,* pp. 13-22, March 2003.

[8] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. IEEE,* vol. 86, no. 11, pp. 2210-2239, Nov. 1998.

[9] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory,* vol. 19, no. 4, pp. 471-480, July 1973.

[10] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. on Information Theory,* vol. 39, no. 3, pp. 821-834, May 1993.

[11] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Information Theory,* vol. 22, no. 1, pp. 1-10, Jan. 1976.

[12] R. Zamir and S. Shamai, "Nested linear/lattice codes for Wyner-Ziv encoding," *Proc. IEEE Information Theory Workshop,* pp. 92-93, June 1998.