JB 96-01

Entitled:

A Statistical Complexity Framework for Topology Preserving Adaptive Vector Quantization

Authors:

M.K. Sonmez, and J.S. Baras

Conference

CISS '96 Conference
March 20-22, 1996
Princeton, NJ

A STATISTICAL MODEL COMPLEXITY FRAMEWORK FOR TOPOLOGY PRESERVING ADAPTIVE VECTOR QUANTIZATION

M. Kemal Sönmez *John S. Baras*

Institute for Systems Research and Electrical Engineering Department, University of Maryland College Park, MD 20742

ABSTRACT

We propose a statistical model complexity framework for topology preserving adaptive vector quantization. In this setting, adaptation of the neighborhood function during training of the codebooks, which is essential for producing global organization, may be regarded as increasing the statistical model complexity as more data become available. Therefore, the training is equivalent to on the fly optimization of the bias/variance trade-off.

1. INTRODUCTION

We are interested in statistical models which adapt their complexity for bias/variance trade-off on-line. Such schemes are useful in applications where estimation and usage of the model for inference must be carried out simultaneously on the fly. Rather than selecting a fixed model size k for a given data set of size n, we would like to have a model with adaptive complexity which starts from a very simple model when very few data are available, then increases its complexity as more data arrive thereby optimizing the bias/variance trade-off on the fly.

The neighborhood adaptation in topology preserving adaptive vector quantization, which is essential for producing global organization, may, in a suitable framework, be regarded as a model with adaptive complexity. Incremental and sparse variations of generalized EM algorithms, when viewed in a free energy minimization context, provide a setting where this kind of adaptive complexity may be justified in terms of maximum likelihood.

2. TOPOLOGICALLY CONSTRAINED ADAPTIVE VQ

Data from a feature space $\mathcal{F}$ is compressed by a VQ codebook, $\mathbf{X}^h = \{\mathbf{x}_k^h \in \mathcal{F}, k = 1, \ldots, K\}$ where each codevector $\mathbf{x}_k^h$ in the feature space $\mathcal{F}$ represents a class of feature vectors. Expressed in terms of distortion, the topologically constrained VQ is the minimization of the modified distortion [1]

$$D' = E[d(\mathbf{x}, \mathbf{x}_w^h)] + \sum_{i \neq w} n_{wi} E[d(\mathbf{x}, \mathbf{x}_i^h)] \qquad (1)$$

where the "winner" w is $w = \arg\min_j |\mathbf{x}(t) - \mathbf{x}_j^h(t)|^2$.

The topology of a reference environment, i.e. the local neighborhood relations, is captured with the neighborhood

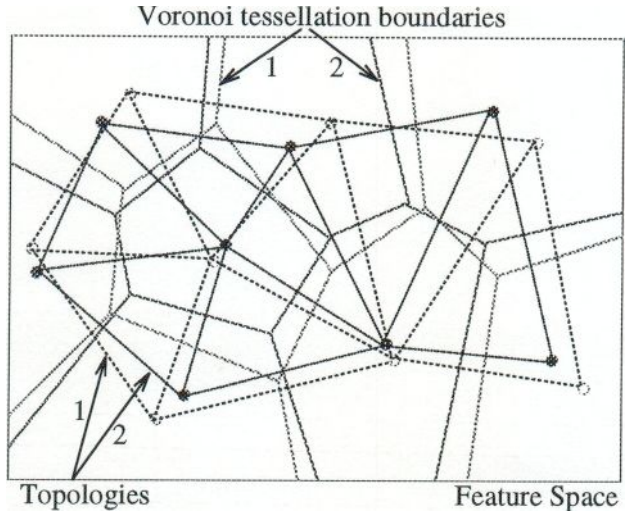

Figure 1. VQ codebooks of two environments (1 and 2) with the topology preserved.

function

$$n_{ij} = \exp\left(-\frac{|\mathbf{x}_i^{ref} - \mathbf{x}_j^{ref}|^2}{2\sigma^2}\right) \forall i, j = 1, \ldots, K \qquad (2)$$

where the VQ codebook for the reference environment, $\mathbf{X}^{ref} = \{\mathbf{x}_k^{ref} \in \mathcal{F}, k = 1, \ldots, K\}$ is designed using the Generalized Lloyd algorithm [2]. The design criterion is the minimization of the distortion

$$D = E[d(\mathbf{x}, \mathbf{x}_w^{ref})] \qquad (3)$$

where the "winner", w, is given by

$$w = \arg\min_j |\mathbf{x} - \mathbf{x}_j^{ref}|^2 \qquad (4)$$

Practically, only N-closest neighbors (N=5-10) are kept and the rest of the n_{ij}'s are set to zero leading to the representation of the topology as an elastic mesh as in Figure 1.

The idea of topology preservation is detailed in [1]. It is one of the mathematical frameworks for the Self Organizing Map (SOM), a topology preserving adaptive VQ algorithm. The topology of SOM is on a non-linear projection of the

signal space onto a 1D or 2D "map", and is arbitrarily decided a priori. The topology may also directly be on the feature space and be simply determined by the distortion measure in the reference environment as explained above.

The minimization of D' is accomplished by the Robbins-Monro stochastic approximation technique, which in the case of squared error distortion reduces to the incremental adaptation

$$\mathbf{x}_k^h(t+1) = \mathbf{x}_k^h(t) + n_{wk}(t)[\mathbf{x}(t) - \mathbf{x}_k^h(t)] \qquad (5)$$

where $\mathbf{x}(t)$, $t = 0, \ldots, T$ are the data from the adaptation environment whose codebook is initialized with the codebook of the reference environment

$$\mathbf{x}_k^h(0) = \mathbf{x}_k^{ref}, \quad k = 1, \ldots, K \qquad (6)$$

or simply set to random vectors in the feature space as in SOM. Notice that in the adaptation, the neighborhood function is also dynamic in the following form

$$n_{ij}(t) = \alpha(t) \exp\left(-\frac{|\mathbf{x}_i^{ref} - \mathbf{x}_j^{ref}|^2}{2\sigma^2(t)}\right) \forall i, j = 1, \ldots, K. \quad (7)$$

The codebook adaptation amounts to stretching of the mesh to better fit the new environment while keeping the neighborhood relations intact.

The functions $\alpha(t)$ and $\sigma^2(t)$ are monotonically decreasing. For initial large values of $\sigma^2(t)$, all codevectors are updated similarly, thus the algorithm in the beginning may be regarded as an incremental version of simple mean normalization. The scale of adaptation is made finer by decreasing $\sigma^2(t)$ as increasingly more data are used.

3. ELASTICITY AND COMPLEXITY

Consider the following analogy for the role of the neighborhood function: For $\sigma^2(t) \to \infty$, we have a rigid mesh in which every codevector gets updated in the same manner. In the case $\sigma^2(t) = 0$, we really have no mesh, only the winner gets updated (unconstrained VQ). For the case where $0 < \sigma^2(t) < \infty$, we have an equivalent of an elastic mesh where, depending on the winner, every codevector gets updated differently. Neighborhood function is also effectively the probability distribution of the winner, $\tilde{P}$, with complexity measured by its entropy as in Table 1. Therefore, the

Neighborhood	K_{eff}	Complexity	Entropy
uniform	1	least	$\log(K)$
finite σ	$K2^{-H}$	↕	$0 < H < \log(K)$
delta	K	most	0

more elastic a mesh associated with a codebook is, the more complex is the model (i.e. has a larger effective number of parameters). The neighborhood function's radius, $\sigma^2(t)$, determines the instantaneous complexity which starts from a single location parameter and increases up to the size of the codebook.

Model selection criteria such as [4]

$$MDL(k) = -\log(P(z|\hat{\theta})) + \frac{k}{2}\log n$$

and

$$AIC(k) = -2\log(P(z|\hat{\theta})) + 2k$$

yield a fixed optimal complexity for the given data batch. Rather than selecting a model size for a given data set of size n, we would like to have a complexity adaptation which starts from a very simple model when very few data are available, then increases its complexity as more data arrive thereby optimizing on the run the bias/variance trade-off. The elasticity discussion suggests that the entropy of the neighborhood function, i.e. the distribution of the winner might be a good indication of instantaneous complexity.

4. INCREMENTAL AND SPARSE EM ALGORITHMS

This idea may be justified in an EM framework via maximization of negative "free energy": [3]

$$F(\tilde{P}, \theta) = E_{\tilde{P}(t)}[\log P(y, z|\theta)] + H(\tilde{P})$$

where random variables Y, Z are the observed and the unobserved variables respectively. The problem is to find the maximum likelihood estimate for the parameters of a model for Y and Z. The main result is: If F has a local maximum at $(\tilde{P}^*, \theta^*)$, then the likelihood also has a local maximum at θ^*. Both E and M steps can be regarded as maximizing a single joint function of the model variables and the distribution of the unobserved variables. Therefore, any suitable optimization scheme that will maximize F will produce a MLE for the incomplete observation problem. This framework allows for incremental and sparse variations of generalized EM algorithms with provable convergence properties.

Generalized Lloyd algorithm, (LBG, k-means) may be regarded as an winner-take-all (i.e. $\tilde{P}^{(t)}(y)$ is 1 for a single value of y and 0 for others) version of the EM algorithm as applied to the Gaussian mixture problem with mixing proportions and variances fixed. In a similar manner, topology preserving adaptive VQ can be regarded as an incremental *and* sparse version of the EM algorithm as applied to the Gaussian mixture problem with mixing proportions and variances fixed. The increase in F at each incremental iteration, therefore, justifies the adaptation of the neighborhood function as on-line complexity optimization. This framework may also be used to find good annealing schedules when the underlying distribution is known to belong to a given class of distributions.

REFERENCES

[1] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin 1995

[2] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, January 1980.

[3] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants". submitted to *Biometrika*. 1993.

[4] J. Rissanen, *Stochastic complexity in statistical inquiry*, World Scientific, Teaneck, N.J. 1989.