ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

On the Dynamics of Vector Quantization and Neural Gas

Aree Witoelar[1], Michael Biehl[1], Anarta Ghosh[1] and Barbara Hammer[2]

1- University of Groningen - Mathematics and Computing Science
P.O. Box 800, NL-9700 AV Groningen - The Netherlands

2- Clausthal University of Technology - Institute of Computer Science
D-98678 Clausthal-Zellerfeld - Germany

Abstract. A large variety of machine learning models which aim at vector quantization have been proposed. However, only very preliminary rigorous mathematical analysis concerning their learning behavior such as convergence speed, robustness with respect to initialization, etc. exists. In this paper, we use the theory of on-line learning for an exact mathematical description of the training dynamics of Vector Quantization mechanisms in model situations. We study update rules including the basic Winner-Takes-All mechanism and the Rank-Based update of the popular Neural Gas network. We investigate a model with three competing prototypes trained from a mixture of Gaussian clusters and compare performances in terms of dynamics, sensitivity to initial conditions and asymptotic results. We demonstrate that rank-based Neural Gas achieves both robustness to initial conditions and best asymptotic quantization error.

1 Introduction

Vector quantization (VQ) is an important unsupervised learning algorithm, widely used in different areas such as data mining, medical analysis, image compression, and speech or handwriting recognition [1]. The main objective of VQ is to represent the data points by a small number of prototypes. This can directly be used for compression, clustering, data mining, or (after post labeling) classification. The basic "winner-takes-all" (WTA) algorithm or batch variants thereof such as the popular k-means clustering directly optimize the quantization error underlying vector quantization. However, since the quantization error is multimodal, these methods can be subject to confinement in local minima and can produce suboptimal results. A variety of alternatives have been proposed, some of which are heuristically motivated, some of which are based on a cost function related to the quantization error: the self-organizing map [9], fuzzy-k-means [2], stochastic optimization [6], or neural gas [10], to name just a few. These algorithms have in common that a pattern influences more than one prototype at a time by using the "winner-takes-most" paradigm. In practice, this often yields better solutions, however, the effect of this strategy on the convergence speed or asymptotic behavior has hardly been rigorously investigated so far.

Methods from statistical physics and the theory of on-line learning [7] allow for an exact mathematical description of learning systems for high dimensional data. In the limit of infinite dimensionality, the system can be fully described in terms of few characteristic quantities. The evolution of these quantities or order parameters along the training procedure can be analysed by a set of coupled ordinary differential equations (ODE). By integrating these ODEs, it is possible

to analyse the performance of VQ algorithms in terms of stability, sensitivity to initial conditions and achievable quantization error.

We extend the theoretical framework of simple (WTA-based) vector quantization introduced in an earlier work [5] by considering more than two prototypes. Further, we investigate the popular Neural Gas approach (NG) [10] as well as a (computationally better tractable) approximation thereof. This is an important step towards the investigation of general VQ approaches which are based on neighborhood cooperation such as NG or the self-organizing map.

2 Winner-Takes-All and Rank-Based Algorithms

Vector Quantization represents N-dim. input data $\xi^\mu \in I\!R^N$ by a set of prototypes $W = \{\mathbf{w}_i\}_{i=1}^S$ in $I\!R^N$. We assume that input data is generated randomly according to a given density $P(\xi)$ and is presented sequentially during training. Depending on the algorithm, one or more prototypes are updated on-line.

The primary goal of VQ is to find a faithful representation of the data by minimizing the so-called quantization or distortion error

$$E(W) = \int d\xi P(\xi) \sum_{i=1}^{S} d(\xi, \mathbf{w}_i) \prod_{j \neq i} \Theta\left[d(\xi, \mathbf{w}_j) - d(\xi, \mathbf{w}_i)\right] - \frac{1}{2} \int d\xi P(\xi) \xi^2 \quad (1)$$

Here we restrict ourselves to the quadratic Euclidean distance measure $d(\xi, \mathbf{w}_i) = (\xi - \mathbf{w}_i)^2/2$. For each input vector ξ the closest prototype $\mathbf{w}_i$ is singled out by the product of Heaviside functions, $\Theta(x) = 0$ if $x < 0$; 1 else. The constant $\frac{1}{2}\int d\xi P(\xi)\xi^2$ term is independent of prototype positions and is substracted for convenience.

Algorithms studied in the following can be interpreted as stochastic gradient descent procedures with respect to a cost function $H(W)$ similar to $E(W)$. The generalized form reads $H(W) = \frac{1}{2}\int d\xi P(\xi) \sum_{i=1}^{S} f(r_i)(\xi - \mathbf{w}_i)^2 - \frac{1}{2}\int d\xi P(\xi)\xi^2$ where r_i is the rank of prototype $\mathbf{w}_i$ with respect to the distance $d(\xi, \mathbf{w}_i)$, i.e. $r_i = S - \sum_{j \neq i} \Theta\left[d(\xi, \mathbf{w}_j) - d(\xi, \mathbf{w}_i)\right]$. Rank $r_J = 1$ corresponds to the so-called *winner*, i.e. the prototype $\mathbf{w}_J$ closest to the example ξ. The rank function $f(r_i)$ determines the update strength for the set of prototypes and satisfies the normalization $\sum_{i=1}^{S} f(r_i) = 1$; note that it does not depend explicitly on distances but only on their ordering.

The corresponding stochastic gradient descent in $H(W)$ is of the form

$$\mathbf{w}_i^\mu = \mathbf{w}_i^{\mu-1} + \Delta\mathbf{w}_i^\mu \quad \text{with} \quad \Delta\mathbf{w}_i^\mu = \frac{\eta}{N} f(r_i)(\xi^\mu - \mathbf{w}_i^{\mu-1}) \quad (2)$$

where η is the learning rate and ξ^μ is a single example drawn independently at time step μ of the sequential training process. We discuss three basic algorithms:

WTA: Only the winner is updated for each input. The corresponding rank function is $f_{\mathrm{WTA}}(r_i) = 1$ if $r_i = 1$; 0 else. Note that for this choice $H(W)$ reduces to the quantization error $E(W)$.

Linear rank: As a particularly simple rank-dependent update we consider a scheme where the update strength decreases linearly with the rank, $f_{\mathrm{LIN}}(r_i) = (1/C)(S - r_i + 1)$ where $C = \frac{1}{2}S(S+1)$ is a normalization factor.

Neural Gas: The update strength decays exponentially with the rank controlled by a parameter λ. The rank function is $f_{\mathrm{NG}}(r_i) = (1/C(\lambda))h_\lambda(r_i)$ with $h_\lambda = \exp(-r_i/\lambda(t))$ and $C(\lambda) = \sum_{i=1}^{S} \exp(-i/\lambda(t))$. The parameter λ is frequently set large initially and decreased in the course of training. Note that for $\lambda \to 0$ NG becomes identical with WTA training.

3 Model Data

We choose the model distribution of the data to be a mixture of two spherical Gaussian clusters: $P(\xi) = \sum_{\sigma=1,2} p_\sigma P(\xi|\sigma)$ with $P(\xi|\sigma) = \frac{1}{(\sqrt{2\pi})^N} \exp[-\frac{1}{2}(\xi - \ell\mathbf{B}_\sigma)^2]$ where p_σ are the prior probabilities. The components of vectors ξ are unit variance random numbers. The cluster mean vectors are $\ell\mathbf{B}_1$ and $\ell\mathbf{B}_2$ where ℓ controls the separation of cluster centers. $\mathbf{B}_\sigma$ are orthonormal, i.e. $\mathbf{B}_i \cdot \mathbf{B}_j = \delta_{i,j}$ where δ is the Kronecker delta. Note that the clusters strongly overlap and the separation is only apparent in the two-dimensional subspace spanned by $\mathbf{B}_{1,2}$. Hence, it is a non-trivial task to detect the structure in N dimensions.

4 Analysis of the Learning Dynamics

We give a brief description of the theoretical framework and refer to [3, 12] for further details. Following the lines of the theory of on-line learning, e.g. [7], the system can be fully described in terms of a few so-called order parameters in the thermodynamic limit $N \to \infty$. A suitable set of characteristic quantities for the considered learning model is: $R_{i\sigma}^{\mu} = \mathbf{w}_i^{\mu} \cdot \mathbf{B}_\sigma$ and $Q_{ij}^{\mu} = \mathbf{w}_i^{\mu} \cdot \mathbf{w}_j^{\mu}$. Note that $R_{i\sigma}$ are the projections of prototype vectors $\mathbf{w}_i^{\mu}$ on the center vectors $\mathbf{B}_\sigma$ and Q_{ij}^{μ} correspond to the self- and cross- overlaps of the prototype vectors.

From the generic update rule defined above, Eq. (2), we can derive the following recursions in terms of the order parameters:

$$\begin{aligned} N(R_{i\sigma}^{\mu} - R_{i\sigma}^{\mu-1}) &= \eta f(r_i)\big(b_\sigma^{\mu} - R_{i\sigma}^{\mu-1}\big) \\ N(Q_{ij}^{\mu} - Q_{ij}^{\mu-1}) &= \eta\big[f(r_j)\big(h_i^{\mu} - Q_{ij}^{\mu-1}\big) + f(r_i)\big(h_j^{\mu} - Q_{ij}^{\mu-1}\big)\big] + \\ &\quad \eta^2 f(r_i) \times f(r_j) + \mathcal{O}(1/N) \end{aligned} \tag{3}$$

where h_i^{μ} and b_i^{μ} are the projections of the input data vector ξ^{μ}: $h_i^{\mu} = \mathbf{w}_i^{\mu-1} \cdot \xi^{\mu}$, $b_\sigma^{\mu} = \mathbf{B}_\sigma \cdot \xi^{\mu}$. For large N, the $\mathcal{O}(1/N)$ term can be neglected. In the limit $N \to \infty$, the order parameters *self average* [11] with respect to the random sequence of examples. This means that fluctuations of the order parameters vanish and the system dynamics can be described exactly in terms of their mean values. Also for $N \to \infty$ the rescaled quantity $t \equiv \mu/N$ can be conceived as a continuous time variable. Accordingly, the dynamics can be described by a set of coupled ODE [3, 8] after performing an average over the sequence of input data:

$$\begin{aligned} \frac{dR_{i\sigma}}{dt} &= \eta\big(\langle b_\sigma f(r_i)\rangle - \langle f(r_i)\rangle R_{i\sigma}\big) \\ \frac{dQ_{ij}}{dt} &= \eta\big(\langle h_i f(r_j)\rangle - \langle f(r_j)\rangle Q_{ij} + \langle h_j f(r_i)\rangle - \langle f(r_i)\rangle Q_{ij}\big) + \\ &\quad \eta^2 \langle f(r_i) \times f(r_j)\rangle \end{aligned} \tag{4}$$

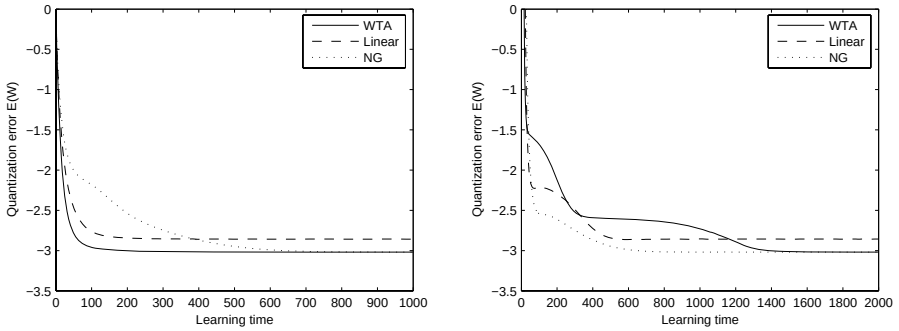


Fig. 1: Evolution of the quantization error $E(W)$ at learning time $\tilde{t} = \eta t$ for WTA, linear rank and NG algorithms for $\eta = 0.1$, $p_1 = 0.75$ and ℓ=1.5. The parameters for NG are $\lambda_i = 2$, $\lambda_f = 0.01$. The set of prototypes is initially set (a) $R_{i\sigma}(0) \approx 0, Q_{ij}(0) \approx 0, \forall\{i,j,\sigma\}$ and (b) $R_{i1}(0) \approx -2, R_{i2}(0) \approx 3, Q_{ij}(0) \approx R_{i1}(0)^2 + R_{i2}(0)^2, \forall\{i,j\}$.

where $\langle . \rangle$ is the average over the density $P(\xi)$.

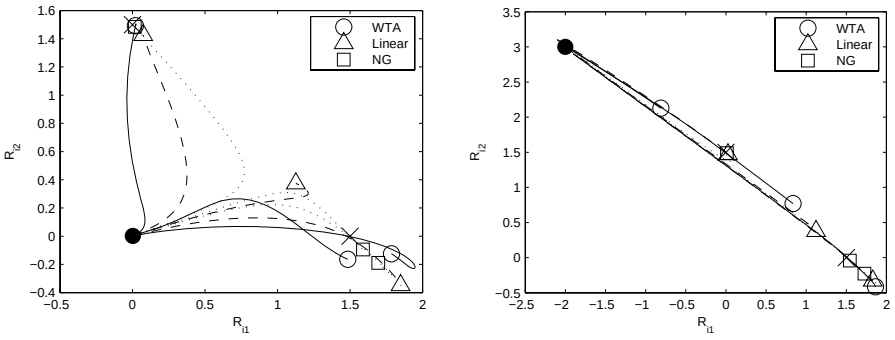
Exploiting the limit $N \to \infty$ once more, the quantities h_i^μ, b_σ^μ become correlated Gaussian quantities by means of the Central Limit Theorem. Thus, the above averages reduce to Gaussian integrations in, here, five dimensions. While most of these can be performed analytically (in particular the linear approximation of NG), some (in particular WTA and NG) have to be implemented numerically. See [3, 12] for details of the computations. Given the averages for a specific rank function $f(r_i)$ we obtain a closed set of ODE. Using initial conditions $R_{i\sigma}(0), Q_{ij}(0)$, we integrate this system for a given algorithm and get the evolution of order parameters in the course of training, $R_{i\sigma}(t), Q_{ij}(t)$.

Analogously, the quantization error, Eq. (1), can be expressed in terms of order parameters after performing the averages in $E = \sum_{i=1}^{S} \big(f_{\mathrm{WTA}}(r_i) Q_{ii} - 2h_i f(r_i) \big)$. Plugging in the values of the order parameters computed by solving the ODE, $\{R_{i\sigma}(t), Q_{ij}(t)\}$, we can study the so called learning curve E in dependence of the training time t for a given VQ algorithm.

5 Results

The dynamics of WTA learning for two prototypes have been studied in an earlier publication [4]. Here we present the non-trivial extension to three competing prototypes and winner-takes-most schemes. The NG algorithm is studied for decreasing λ with $\lambda(t) = \lambda_i(\lambda_f/\lambda_i)^{t/t_f}$ where t_f is the maximum learning time.

As shown in Fig. 1, we observe that the quantization error decreases faster in the WTA algorithm compared to rank-based methods at the initial stages of the learning. This behavior can be explained by the fact that the cost function of winner-takes-most algorithms differs from the quantization error by smoothing terms in particular in early stages of training. WTA yields better asymptotic quantization error than the linear rank which is due to the fact that the approxi-



(a) **(b)**

Fig. 2: The trajectories of the 3 prototypes projected on the plane spanned by the two cluster mean vectors R_{i1}-R_{i2}, using the parameters in Figs. 1(a) and 1(b). The crosses mark the two Gaussian cluster centers $\ell\mathbf{B}_1$ ($p_1 = 0.75$) and $\ell\mathbf{B}_2$. The dots mark the initial positions and the markers ◯, △ and □ indicate the projections of prototypes after $\tilde{t}$=50 for different algorithms.

mation of the exponential of NG by the linear rank differs from the quantization error also in late stages of training. For small λ, WTA and NG training become identical. This is mirrored by the fact that, for large t and $\lambda_f \to 0$, both algorithms yield the same quantization error. The linear approximation of NG describes the behavior of NG quite well in early stages of training, such that it constitutes a feasible approximation for these stages (note that the linear approximation, unlike NG, does not require numerical evaluation of the ODEs).

The influence of initial prototype positions is of particular interest since winner-takes-most schemes are supposed to partially overcome the sensitivity of WTA schemes to initialization. The set of initial values $\{R_{i\sigma}(0), Q_{ij}(0)\}$ strongly affects the later performance of the algorithms. In Fig. 1(a), prototypes are initialized close to the origin and we observe that WTA yields the best overall quantization error in this case. In Fig. 1(b), we set the initial prototypes on the extension of the line between cluster centers, viz. $\ell(\mathbf{B}_1 - \mathbf{B}_2)$. Here, all prototypes were initialized far away from the origin on the side of the weaker cluster. For WTA training, prototypes reach $t \to \infty$ asymptotic positions corresponding to the global minimum of $E(W)$ for small learning rates $\eta \to 0$. However, learning can slow down significantly at intermediate stages of the training process. Transient configurations may persist in the vicinity of local minima and can indeed dominate the training process. Rank-based methods are more robust w.r.t. the initial position of prototypes than WTA. Apparently, the NG combines the advantages of robustness to initial conditions with achieving the best quantization error asymptotically.

The projections of the prototype vectors on the plane spanned by the cluster centers $\mathbf{B}_1$ and $\mathbf{B}_2$ after $\tilde{t} = 50$ are presented in Fig. 2. In WTA, the projections of two prototypes converge near the center of the stronger Gaussian cluster. The two prototypes, say i and j, have the same $R_{i\sigma} = R_{j\sigma}$ and length $Q_{ii} = Q_{jj}$. However, the prototypes differ in their components orthogonal to the $\mathbf{B}_1 - \mathbf{B}_2$ plane. Therefore, they are not identical vectors, as they do not satisfy the

condition $Q_{ii} = Q_{jj} = Q_{ij} \iff \mathbf{w}_i = \mathbf{w}_j$. It is also worth noting that for all algorithms, the projections of the asymptotic configuration are located on the $\mathbf{B}_1 - \mathbf{B}_2$ line due to symmetry reasons.

6 Conclusion

We have put forward an exact mathematical analysis of the dynamics of WTA- and rank-based VQ algorithms for three prototypes in a high dimensional data space. The performance is measured by the evolution of the quantization error. The WTA algorithm always converges to the best asymptotic quantization error in this comparably simple learning scenario, however the learning curve is highly dependent on the initial conditions. The rank-based methods are less sensitive to the initial conditions, and NG in particular achieves both robustness and best asymptotic quantization error. Thereby, convergence speed is comparable or (for initialization outside the clusters) better than simple WTA mechanisms and the same final quantization error can be obtained. Since NG cannot fully be solved analytically but requires numerical integration of the corresponding ODEs, we also investigated a simple linear approximation. This reasonably approximates the quantization error of NG during early stages of training, but performs worse in the asymptotic limit since the cost function is very different from $E(W)$.

References

[1] *Bibliography on the Self Organising Map (SOM) and Learning Vector Quantization (LVQ)*, Neural Networks Research Centre, Helsinki University of Technology, 2002.

[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[3] M. Biehl, A. Freking, A. Ghosh and G. Reents, *A theoretical framework for analysing the dynamics of LVQ*, Technical Report, Technical Report 2004-09-02, Mathematics and Computing Science, University Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands, December 2004, available from *www.cs.rug.nl/∼biehl*.

[4] M. Biehl, A. Freking and G. Reents, Dynamics of On-line Competitive Learning *Europhysics Letters*, 38: 73-78, 1996.

[5] M. Biehl, A. Ghosh and B. Hammer, Learning Vector Quantization: The Dynamics of Winner-Takes-All Algorithms. *Neurocomputing*, 69: 660-670, 2006.

[6] J. Buhmann, Stochastic Algorithms for Exploratory Data Analysis: Data Clustering and Data Visualization, in *Learning in Graphical Models*, M. Jordan (ed.), Kluwer, 1997.

[7] A. Engel and C. van Broeck, editors. *The Statistical Mechanics of Learning*, Cambridge University Press, 2001

[8] A. Ghosh, M. Biehl and B. Hammer, Performance Analysis of LVQ Algorithms: A Statistical Physics Approach, *Neural Networks*, special issue on Advances in Self-Organizing Maps. Vol. 19:817-829, 2006.

[9] T. Kohonen. *Self Organising Maps*, Springer, 3rd ed., 2001

[10] T. Martinetz, S. Berkovich, K. Schulten, 'Neural Gas' network for vector quantization and its application to time series prediction, *IEEE TNN*, 4(4):558-569, 1993.

[11] G. Reents and R. Urbanczik, Self Averaging and On-line Learning, *Phys. Rev. Letter*, 80:5445-5448, 1998.

[12] A. Witoelar and M. Biehl, *Dynamics of multiple prototype VQ*, Technical Report, in progress, Mathematics and Computing Science, University Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands, December 2006