ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Phase transitions in Vector Quantization

Aree Witoelar[1], Anarta Ghosh[2], Michael Biehl[1]

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

2- Nordic Bioscience - Imaging Division
Herlev Hovedgade 207, 2730 Herlev - Denmark

Abstract. We study Winner-Takes-All and rank based Vector Quantization along the lines of the statistical physics of off-line learning. Typical behavior of the system is obtained within a model where high-dimensional training data are drawn from a mixture of Gaussians. The analysis becomes exact in the simplifying limit of high training temperature. Our main findings concern the existence of phase transitions, i.e. a critical or discontinuous dependence of VQ performance on the training set size. We show how the nature and properties of the transition depend on the number of prototypes and the control parameter of rank based cost functions.

1 Introduction

Vector Quantization (VQ) is one of the most important families of algorithms for unsupervised learning. It has been applied in a large variety of practical contexts, see [1] for examples and references. The aim of VQ is the faithful representation of a large amount of data by only a few prototype vectors, thus detecting structures that are present in the data.

Competitive learning algorithms such as Winner-Takes-All (WTA) schemes or batch variants like the popular k-means clustering aim directly at the minimization of the quantization error. However, they may suffer from confinement in local minima, potentially leading to far from optimal performance. Numerous modifications have been suggested in order to overcome this difficulty. Prominent examples are Self-Organizing Maps [7], fuzzy k-means [3], or Neural Gas (NG) [9], to name only a few. They have in common that the WTA prescription is replaced by schemes which assign each data point to more than one prototype. In particular, NG algorithms employ rank based cost functions [9].

We analyse and compare WTA and rank based training within a model scenario. In previous studies we addressed the dynamics of on-line VQ and NG schemes which are based on a sequence of single example data, e.g. [4, 13]. Here, we consider training from a set of examples by means of off-line or batch stochastic optimization of a cost function. To this end, we apply methods from the equilibrium physics of learning which were formerly used to study, amongst others, multilayer neural networks [6, 10, 12]. This approach allows us to investigate the typical behavior of off-line VQ learning schemes in model situations.

Our analysis is based on the simplifying limit of training at high temperatures which has proven to yield first insights into many training scenarios [5, 6, 10]. It shows how invariances with respect to permutation of prototypes lead to phase transitions which govern the training process: A critical number of examples is required for the successful detection of the underlying structure. Similar effects

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

of "retarded learning" have been studied in several models and learning scenarios earlier, e.g. [5, 8, 11]. Here we consider the extensions to rank based training and scenarios with more than two prototypes. We show that the nature of the transition can change significantly under these modifications.

2 Vector Quantization Cost Functions

Assume a data set of P examples is given as $I\!D = \{\boldsymbol{\xi}^\mu \in I\!R^N\}_{\mu=1}^P$. We consider a system of K prototype vectors $\mathbf{W} = \{\mathbf{w}_k \in I\!R^N\}_{k=1}^K$ with $K \ll P$. The cost functions considered here can be expressed as empirical averages of an error measure:

$$H(\mathbf{W}) = \sum_{\mu=1}^P e(\mathbf{W}, \boldsymbol{\xi}^\mu) \;\text{ with }\; e(\mathbf{W}, \boldsymbol{\xi}) = \tfrac{1}{2} \sum_{k=1}^K d(\mathbf{w}_k, \boldsymbol{\xi})\, g(r_k) - \tfrac{1}{2}\boldsymbol{\xi}^2. \quad (1)$$

Here the last term is constant w.r.t. the choice of $\mathbf{W}$ and is subtracted for convenience in later calculations. Throughout the following, we employ the squared Euclidean distance measure $d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^2$. In Eq. (1), the normalization $\sum_{k=1}^K g(r_k) = 1$ of the so-called rank function g is assumed. The argument r_k is the rank of prototype $\mathbf{w}_k$ with respect to its distance from input vector $\boldsymbol{\xi}$. It can be written as

$$r_k = K - \sum_{j \neq k}^K \Theta_{kj} \quad \text{with the shorthand} \quad \Theta_{kj} = \Theta\left[d(\boldsymbol{\xi}, \mathbf{w}_j) - d(\boldsymbol{\xi}, \mathbf{w}_k)\right] \quad (2)$$

where $\Theta(.)$ is the Heaviside function. Specifically, we consider rank functions of the form

$$g_\lambda(r_i) = \exp[-r_i/\lambda] \Big/ \sum_{k=1}^K \exp[-r_k/\lambda], \quad (3)$$

where λ controls the soft assignment of a given vector $\boldsymbol{\xi}$ to the prototypes. In the limit $\lambda \to 0$ only the winner $\mathbf{w}_J$ with $r_J = 1$ is taken into account, $g_o(k) = \delta_{k,1}$, and the costs, Eq. (1), reduce to the quantization error with

$$e_{VQ}(\mathbf{W}, \xi) = \tfrac{1}{2} \sum_{i=1}^K d(\mathbf{w}_i, \boldsymbol{\xi}) \prod_{j \neq i}^K \Theta_{ij} - \tfrac{1}{2}\boldsymbol{\xi}^2. \quad (4)$$

Note that the cost functions considered here are invariant under exchange or permutations of prototypes.

3 Model Data

We study training processes where the examples $\boldsymbol{\xi}^\mu$ are generated independently according to a given model density. We will exploit the thermodynamic limit $N \to \infty$ and assume that the number of examples also grows linearly in N, i.e. $P \propto N$. Specifically, we consider a mixture of two spherical Gaussian clusters:

$$P(\boldsymbol{\xi}) = \sum_{m=1}^2 p_m\, P(\boldsymbol{\xi}|m) \text{ with } P(\boldsymbol{\xi}|m) = \tfrac{1}{(2\pi)^{N/2}} \exp\left[-\left(\boldsymbol{\xi} - \ell \mathbf{B}_m\right)^2/2\right] \quad (5)$$

where the prior weights satisfy $p_1 + p_2 = 1$. The cluster centers are given by $\ell\mathbf{B}_1$ and $\ell\mathbf{B}_2$ with the separation parameter ℓ. Without loss of generality, we assume that the $\mathbf{B}_m$ are orthonormal with $\mathbf{B}_m \cdot \mathbf{B}_n = \delta_{mn}$. Densities of the above or a

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

similar form have been studied previously in the context of both supervised and unsupervised learning, see e.g. [4, 6, 8, 13]. Note that for large N, the highly overlapping clusters become only apparent in subspaces that have significant overlap with the $\mathbf{B}_i$. Projections into randomly selected two-dimensional spaces, for instance, do not display any structure, see [4].

4 Equilibrium Physics Approach

We give a brief overview of the standard statistical physics analysis of off-line learning [10, 12] and refer to [14] for the details. Training is interpreted as a stochastic minimization of $H(\mathbf{W})$ on the data set $I\!D$, where the formal temperature T controls the degree of randomness. This leads to a well-defined thermal equilibrium: a configuration $\mathbf{W}$ is observed with a probability given by the Gibbs density

$$P(\mathbf{W}) \;=\; \exp\left[-\beta\, H(\mathbf{W})\right]/\, Z \;\;\text{where}\;\; Z \;=\; \int d\mu(\mathbf{W})\, \exp\left[-\beta\, H(\mathbf{W})\right]. \qquad (6)$$

Here $\beta = 1/T$, the normalization Z is called the partition sum and the measure $d\mu(\mathbf{W})$ is the NK-dim. volume element. Thermal averages $\langle . \rangle$ over $P(\mathbf{W})$ can be calculated as derivatives of the so-called free energy $-\ln Z/\beta$, for instance: $\langle H \rangle = -\,\partial \ln Z / \,\partial\beta$.

Note that this type of average describes the system trained on one specific data set. In order to obtain generic properties of the model scenario, an additional average over all possible $I\!D$ is performed, yielding the so-called quenched free energy $-\left\langle \ln Z \right\rangle_{I\!D} / \beta$ [6, 10, 12]. In general, the computation of $\left\langle \ln Z \right\rangle_{I\!D}$ requires involved techniques from the theory of disordered systems such as the replica method.

Here we resort to the study of training at high temperatures which allows us to use simplifying relations in the limit $\beta \to 0$. This limit has proven to provide important insights into a variety of learning scenarios [6, 10, 12]. Non-trivial results can only be expected if the increased noise is compensated for by a larger number of examples $\widetilde{\alpha} = \beta(P/N)$. Because large training sets sample the model density very well, the empirical average $\frac{1}{P}\sum_{\mu}^{P} e(\mathbf{W}, \boldsymbol{\xi}^{\mu})$ can be replaced by $\langle e \rangle_{\xi}$, i.e. an average over the full $P(\boldsymbol{\xi})$.

The mean cost $\langle e \rangle_{\xi}$ for high dimensional data can be expressed as a function of the order parameters

$$R_{ij} \;=\mathbf{w}_i \cdot \mathbf{B}_j \;\;\text{and}\;\; Q_{ij} \;=\mathbf{w}_i \cdot \mathbf{w}_j, \qquad (7)$$

see [13] for the result and details of the calculation. It can be performed analytically for systems with two prototypes and involves numerical Gaussian integrals for $K \geq 3$. The set of quantities (7) represents the structure imposed by the cluster center vectors $\mathbf{B}_j$. We can rewrite $\left\langle \ln Z \right\rangle_{I\!D}$ as an integral over the order parameters as follows:

$$\left\langle \ln Z \right\rangle_{I\!D} = \ln \int \textstyle\prod_{i,j} dR_{ij} \prod_{i,j\leq i} dQ_{ij} \exp\left(-N \left[\widetilde{\alpha}\, \langle e \rangle_{\xi} \;-\; s\left(\{R_{ij}, Q_{ij}\}\right)\right]\right). \qquad (8)$$

Here, the entropy term s gives the phase space volume corresponding to a particular configuration of order parameters $\{R_{ij}, Q_{ij}\}$, see [2, 14] for a derivation and the result in closed form.

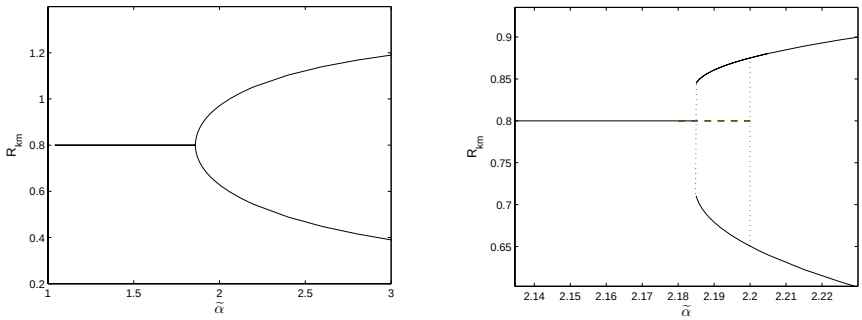

Fig. 1: (a) The order parameters R_{k1} of the stable configuration given the number of example $\widetilde{\alpha}$ for $K=2$. The system undergoes a continuous phase transition at a critical value $\widetilde{\alpha}_c(K=2)\approx 1.85$. (b) R_{k1} for $K=3$, with two of the three values coinciding in the upper curve. The transition is discontinuous; solid (dashed) lines mark global (local) minima of f. Here, $\widetilde{\alpha}_s(K=3)\approx\widetilde{\alpha}_c(K=3)\approx 2.18$ and $\widetilde{\alpha}_d(K=3)\approx 2.20$. The parameters of the input density (5) are $p_1=0.8, p_2=0.2$ and $\ell=1$ in both panels.

We can use the saddle-point method to evaluate (8) in the limit of large N. For $N\to\infty$, this integral is dominated by the maximum integrand, i.e. the minimum of the terms in the square brackets $f(\{R_{ij},Q_{ij}\})=\widetilde{\alpha}\,\langle e\rangle_\xi - s\,(\{R_{ij},Q_{ij}\})$, and the quenched free energy becomes $-\langle\ln Z\rangle_{\mathbb{D}}/N=\beta\min f(\{R_{ij},Q_{ij}\})$. Hence, given a specific cost function and training set size $\widetilde{\alpha}$, we obtain the typical equilibrium properties of the system by minimizing the free energy function $f(\{R_{ij},Q_{ij}\})$ with respect to the order parameters.

5 Results

We first investigate the WTA cost function with $\lambda=0$. For $K=2$ and small $\widetilde{\alpha}$, thermal equilibrium corresponds to states with unspecialized prototypes, i.e. the specialization $\Delta_m=|R_{1m}-R_{2m}|=0$ for all m. Both prototypes $\mathbf{w}_{1,2}$ coincide in the space spanned by $\mathbf{B}_{1,2}$, while their differences in the $(N-2)$-dim. orthogonal space are reflected by non-trivial configurations of $\{Q_{ij}\}$. The underlying cluster structure is not at all detected as long as $\widetilde{\alpha}$ is smaller than the critical value $\widetilde{\alpha}_c$. This parallels findings for supervised learning in neural networks with two hidden units [5] or unsupervised learning scenarios [8, 11]. Above $\widetilde{\alpha}_c$, prototypes begin to align with the clusters and the system becomes specialized, i.e. each $\mathbf{w}_i$ has a larger overlap with exactly one of the cluster centers. Obviously, exchange of the prototypes would not alter the value of H or f and the two configurations are completely equivalent. In the continuous symmetry breaking transition, one of the two states is selected as signaled by a sudden power law increase of Δ_m for $\widetilde{\alpha}\geq\widetilde{\alpha}_c$. Fig. 1 (a) shows the dependence of the equilibrium values of R_{11} and R_{21} on $\widetilde{\alpha}$ in an example situation. The transition results in a non-differentiable kink in the learning curve $\langle e_{VQ}\rangle_\xi$ vs. $\widetilde{\alpha}$ as shown in Fig. 2 (a). The critical value

depends on the model settings. For instance, $\widetilde{\alpha}_c$ will be larger for smaller ℓ.

The behavior is qualitatively different in systems with $K = 3$, see Fig. 1 (b). Again, equilibrium configurations are unspecialized for small data sets. At a characteristic value $\widetilde{\alpha}_s$, a specialized configuration with lower $\langle e_{VQ} \rangle_\xi$ appears. However, the transition is discontinuous, i.e. the specialization Δ_m increases from zero to a finite value in $\widetilde{\alpha}_s$. The projections of two of the three prototypes into the span$(\mathbf{B}_1, \mathbf{B}_2)$ coincide close to the center of the cluster with larger prior weight. Note that, in a generic discontinuous phase transition, one expects a range of values $\widetilde{\alpha}_s \leq \widetilde{\alpha} < \widetilde{\alpha}_c$ where the specialized configuration corresponds to a local minimum of f, see [5] for an example. However, for the setting of parameters considered here, $\widetilde{\alpha}_s = \widetilde{\alpha}_c$ within the achievable numerical precision and we find that the free energy of the specialized configuration is always smaller than that of the system with $\Delta_m = 0$. However, a local minimum of f corresponding to unspecialized $\mathbf{w}_i$ persists in the range $\widetilde{\alpha}_c \leq \widetilde{\alpha} < \widetilde{\alpha}_d$. While such a meta-stable state does not represent thermal equilibrium, its existence can have strong delaying effects in the practical optimization of $H(\mathbf{W})$. We expect $\widetilde{\alpha}_s$ and $\widetilde{\alpha}_c$ to differ more significantly for other settings of the model parameters, e.g. for larger ℓ.

Finally we investigate the minimization of rank based cost functions with $\lambda > 0$ in Eq. (1). We observe the same qualitative behavior as in WTA learning. However, the critical value α_c needed for prototype specialization and thus successful training, increases with the rank function parameter λ. Figure 2 (c) shows this dependence for the two example cases with $K = 2$ and $K = 3$. Note that the slope $d\widetilde{\alpha}_c/d\lambda = 0$ for $\lambda \to 0$. Thus, performing rank based training with an appropriate annealing of λ appears to be a promising strategy for practical optimization of the quantization error. In this context, also the dependence of $\widetilde{\alpha}_d$ on λ will play an important role.

6 Conclusion

We have presented first results for WTA and rank based VQ systems along the lines of the statistical physics analysis of off-line learning. The analysis is based on the high temperature limit, which provides important insights into the training process. In analogy to previous studies of supervised learning, compare e.g. [5] with [2], we expect that most of our findings will carry over qualitatively to stochastic minimization procedures at finite temperature, i.e. low noise.

We show that for, both, two- and three-prototype systems, a critical number of examples is required before the underlying structure can be detected at all. This parallels findings for various other training scenarios and is highly relevant from a practical point of view: even the best optimization strategies will fail completely if too few example data are available. The nature of the phase transition is continuous for two prototypes and discontinuous for $K \geq 3$. The meta-stable states for $K \geq 3$ also shows that long delays may happen in practice even if the critical number of examples is exceeded.

In future projects we will investigate systematically the dependence of the learning behavior on the model parameters. For instance, we anticipate the existence of further competing states such as partially symmetric configurations

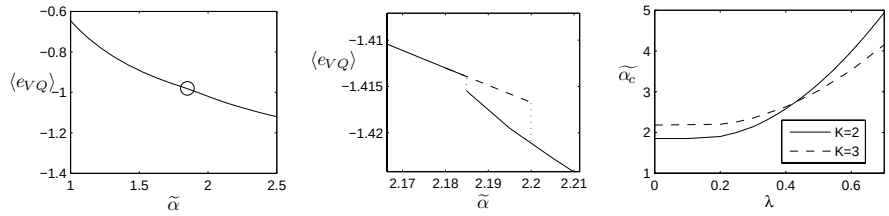

Fig. 2: The mean error $\langle e_{VQ} \rangle_\xi$ for (a) $K=2$ and (b) $K=3$ and parameters as in Fig. 1. The transition results in a kink for $K=2$ and a discontinuous drop for $K=3$ at the respective $\widetilde{\alpha}_c$. c) The values of $\widetilde{\alpha}_c$ vs. the parameter λ, cf. Eq. (3).

for larger ℓ and in NG systems with many prototypes. The analysis of training at low temperature will require more sophisticated techniques, e.g. the so-called annealed approximation or the replica method. These allow for an independent variation of the number of examples P/N and the training temperature and will provide further insight into the typical behavior of practical VQ schemes.

References

[1] *Bibliography on the Self Organising Map (SOM) and Learning Vector Quantization (LVQ)*, Neural Networks Research Centre, Helsinki University of Technology, available from http://liinwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html .

[2] M. Ahr, M. Biehl and R. Urbanczik, Statistical physics and practical training of soft-committee machines, *Eur. Phys. J.* B 10:583, 1999.

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

[4] M. Biehl, A. Ghosh and B. Hammer, Dynamics and generalization ability of LVQ algorithms, *J. Mach. Learning Res.* , 8:323-360, 2007.

[5] M. Biehl, E. Schlösser and M. Ahr, Phase transitions in soft-committee machines, *Europhys. Lett.*, 44(2):261-267, 1998.

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

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

[8] E. Lootens and C. van den Broeck, Analysing cluster formation by replica method, *Europhys. Lett.* 30:381-387, 1995.

[9] 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.

[10] H.S. Seung, H. Sompolinsky and N. Tishby, Statistical mechanics of learning from examples, *Physical Review A*, 45:6056, 1992.

[11] T.L.H. Watkin and J.P. Nadal, Optimal unsupervised learning, *J. Phys. A* 27:1899-1915, 1994.

[12] T.L.H. Watkin, A. Rau and M. Biehl, The statistical mechanics of learning a rule, *Rev. Mod. Phys.*, 65(2):499-556, 1993.

[13] A. Witoelar, M. Biehl, A. Ghosh and B. Hammer, Learning dynamics and robustness of vector quantization and neural gas, *Neurocomputing*, in press.

[14] A. Witoelar, M. Biehl, Equilibrium physics approach in vector quantization. Technical Report, Mathematics and Computing Science, University of Groningen, available from http://www.cs.rug.nl/∼aree.