Phase Transitions in Vector Quantization and Neural Gas

Aree Witoelar [a,*] Michael Biehl [a]

[a] *University of Groningen, Mathematics and Computing Science, P.O. Box 407, NL-9700 AK Groningen, The Netherlands*

Abstract

The statistical physics of off-learning is applied to Winner-Takes-All and rank-based vector quantization (VQ), including the Neural Gas (NG). The analysis is based on the limit of high training temperatures and the annealed approximation. The typical learning behavior is evaluated for systems of two and three prototypes with data drawn from a mixture of high dimensional Gaussian clusters. The learning curves exhibit phase transitions, i.e. a critical or discontinuous dependence of performances on the training set size and training temperature. 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. The NG based systems are demonstrated to give an advantage over WTA in terms of robustness to initial conditions.

Key words: Vector quantization, Clustering, Offline learning, Winner-Takes-All algorithms, Neural Gas

1. Introduction

Vector Quantization (VQ) is an important family of unsupervised learning algorithms and has been applied in many fields, e.g. data mining, image compression and clustering problems [12]. The main objective of VQ is to represent the data faithfully by a small number of prototypes or codebook vectors, measured by the quantization error.

Competitive learning schemes such as the basic "winner-takes-all" (WTA) approach or batch variants such as the k-means algorithm aim at direct minimization of the quantization error. However, such methods are susceptible to confinement in local minima, leading to far from optimal performance. Numerous extensions and modifications have been proposed in order to overcome this difficulty: the self-organizing map (SOM) [9], fuzzy-k-means [2] and neural gas [11], to name just a few. These algorithms have in common that each data point is assigned to more than one prototype. In particular, NG algorithms replace the quantization error by related rank based cost functions [11].

In previous studies we have addressed the dynamics of on-line VQ and NG schemes which are based on a sequence of single example data, e.g. [4,19]. 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 earlier used to study, amongst others, feed-forward neural networks [7,13,17]. The approach allows us to investigate the typical behavior of off-line VQ learning schemes in non-trivial model situations.

Our analysis is based on the so-called annealed approximation which has proven to yield valuable insights into many training scenarios. In particular, it becomes exact in the limit of high training temperatures, which allows for a simplifying description of qualitative behavior. The theory and several applications of annealed approximation and high-temperature limit can be found in, e.g.,

* Corresponding author.
Email address: a.w.witoelar@rug.nl (Aree Witoelar).
URL: http://www.cs.rug.nl/∼aree (Aree Witoelar).

[5,7,13,14,17].

Our analysis of Vector Quantization and Neural Gas shows how invariances with respect to the 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 of "retarded learning" have been studied in several models and learning scenarios earlier, e.g. [5,6,8,10,16].

Extending earlier studies of off-line competitive learning, see [7,10] for an example and further references, we consider rank based training and scenarios with more than two prototypes. We show that the nature of the transition can change significantly under these modifications. 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.

In section 2 we describe the cost functions minimized in the respective VQ learning schemes. This includes the basic WTA and rank-based cost functions. The high-dimensional model data is explained in section 3 while section 4 briefly describes the analysis in the equilibrium physics framework. Sections 5 and 6 present the obtained results for systems with two and three prototypes with emphasis on phase transitions in the learning curves. A summary and outlook is given in section 7.

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:

$$H(\mathbf{W}) = \sum_{\mu=1}^{P} e(\mathbf{W}, \boldsymbol{\xi}^\mu) \;\text{ with}$$

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

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

$$r_k = K - \sum_{j \neq k}^{K} \Theta_{kj} \quad \text{with the shorthand}$$

$$\Theta_{kj} = \Theta\left[d(\boldsymbol{\xi}, \mathbf{w}_j) - d(\boldsymbol{\xi}, \mathbf{w}_k)\right] \quad (2)$$

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

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

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

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

Note that the cost functions considered here are invariant under exchange or permutations of prototypes. This is different from supervised learning where prototypes and data vectors carry class labels. In Learning Vector Quantization [4,9,12], for instance, the permutation symmetry holds only within the classes.

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:

$$P(\boldsymbol{\xi}) = \sum_{m=1}^{2} p_m\, P(\boldsymbol{\xi}|m) \text{ with}$$

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

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 similar form have been studied previously in the context of both supervised and unsupervised learning, see e.g. [4,7,10,20].

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 [13,17] and refer to [18] 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 [7,13,17]

$$F = -\,\langle \ln Z \rangle_{I\!D}\; /\beta. \qquad (7)$$

Proper derivatives thereof yield quantities of interest on average over the randomness contained in $I\!D$ and over the stochastic outcome of the training process. In general, the computation of $\langle \ln Z \rangle_{I\!D}$ requires involved techniques from the theory of disordered systems such as the replica method [7,13,17]

The analysis of thermal equilibrium does not directly correspond to the application of a particular learning algorithm in practical situations. However, it relates to the use of a specific cost function which guides a stochastic training process.

The interpretation of our results can be based on the observation that Eq. (6) corresponds to the stationary density of $\mathbf{W}$ under a so-called Langevin dynamics for well-behaved differentiable energies $H(\mathbf{W})$:

$$\partial \mathbf{W}/\,\partial t = -\nabla_W H(\mathbf{W}) + \mathbf{\Gamma}(t), \qquad (8)$$

see [15,17] for a discussion in the context of learning. Here, $\mathbf{\Gamma}(t)$ is a KN-dim. vector of δ-correlated white noise: $\langle \Gamma_i(t)\Gamma_j(t')\rangle = 2\,T\,\delta_{ij}\,\delta(t-t')$. The approach outlined above yields properties of the stationary density $P(\mathbf{W})$ resulting from (8), on average over the data set contained in $H(\mathbf{W})$.

Practical algorithms will not have precisely the form (8), but one can expect that our results carry over, qualitatively, to more general learning schemes that are guided by the minimization of $H(\mathbf{W})$.

We discuss two important simplifying approaches: the high temperature limit and the annealed approximation.

4.1. *High temperature limit*

First, we study 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 [7,13,17]. Non-trivial results can only be expected if the increased noise is compensated for by a larger number of examples P which scales like

$$P = \widetilde{\alpha}\; (N\,/\,\beta). \qquad (9)$$

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 $P(\boldsymbol{\xi})$. Consequently, training set and test set performances coincide in this simplifying limit. Following the calculations in Appendix A, the averaged logarithm of the partition sum can be rewritten as

$$\langle \ln Z \rangle_{I\!D} = \ln \int d\mu(\mathbf{W}) \exp\Big[-\widetilde{\alpha} N\, \langle e \rangle_{\boldsymbol{\xi}}\Big] \qquad (10)$$

where the rescaled number of examples $\widetilde{\alpha}$ plays the role of an effective inverse temperature and $N\,\langle e \rangle_{\xi}$ is the extensive energy of the system.

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

$$R_{ij} \;= \mathbf{w}_i \cdot \mathbf{B}_j \;\; \text{and} \;\; Q_{ij} \;= \mathbf{w}_i \cdot \mathbf{w}_j, \qquad (11)$$

see [4,20] 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 (11) represents the structure imposed by the cluster center vectors $\mathbf{B}_j$. We can rewrite $\langle \ln Z \rangle_{I\!D}$ as an integral over the order parameters as follows:

$$\langle \ln Z \rangle_{I\!D} = \ln \int \left[\prod_{i,j} dR_{ij} \right] \left[\prod_{k,l \le k} dQ_{kl} \right] \times \exp\left(-N f(\{R_{ij}, Q_{kl}\})\right), \quad (12)$$

where f is called the free energy function,

$$f(\{R_{ij}, Q_{kl}\}) = \widetilde{\alpha} \langle e \rangle_\xi - s(\{R_{ij}, Q_{kl}\}). \quad (13)$$

The right hand side can be obtained as a function of the order parameters in closed form. The entropy term s relates to the phase space volume corresponding to a particular configuration of order parameters $\{R_{ij}, Q_{kl}\}$.

We can use the saddle-point method to evaluate (12) in the limit of large N. For $N \to \infty$, the integral is dominated by the maximum integrand, i.e. the minimum of f. The quenched free energy becomes

$$-\langle \ln Z \rangle_{I\!D} / N = \beta \min f(\{R_{ij}, Q_{kl}\}). \quad (14)$$

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 $f(\{R_{ij}, Q_{kl}\})$ with respect to the order parameters. The corresponding $\{R_{ij}, Q_{kl}\}$ describe the typical properties of the configurations that dominate the Gibbs ensemble.

4.2. *Annealed Approximation*

Practical training procedures aim at an efficient minimization of the cost function. In the statistical physics interpretation of the learning process, this corresponds to low temperatures. While the correct treatment of finite T requires sophisticated techniques such as the replica trick, a useful approximation method which is technically less difficult, can be employed to perform the quenched average $\langle \ln Z \rangle_{I\!D}$.

In the so-called annealed approximation, $\langle \ln Z \rangle_{\boldsymbol{\xi}}$ is approximated by the logarithm of the averaged Z instead. It is equivalent to the approximation

$$\left\langle \frac{\exp(-\beta H(\mathbf{W}))}{Z} \right\rangle_{I\!D} \approx \frac{\langle \exp(-\beta H(\mathbf{W})) \rangle_{I\!D}}{\langle Z \rangle_{I\!D}} \quad (15)$$

The annealed approximation becomes exact in the limit $\beta \to 0$ and coincides with the explicit treatment of this limit [13]. At low temperatures the annealed free energy yields only an upper bound to the correct one, but the hope is that the position of minima in terms of the $\{R_{ij}, Q_{kl}\}$ is similar. The scheme has proven useful in predicting qualitative behavior of many learning systems, e.g. [7,15]. The validity of the annealed approximation is discussed systematically in, for instance, [13,14].

The average partition function can be rewritten as

$$\langle Z \rangle_{I\!D} = \int d\mu(\mathbf{W}) \exp\left[-\alpha N G_A(\mathbf{W})\right] \quad \text{with} \quad G_A = -\ln \langle \exp(-\beta e(\boldsymbol{\xi}^\mu, \mathbf{W})) \rangle_\xi \quad (16)$$

where G_A involves an average over one random input only. Only for $\beta \to 0$ this average can be absorbed into the exponent and we recover the high temperature result.

The calculation of G_A can be done analytically for two prototypes and arbitrary β, as outlined in [18]. The corresponding free energy function as in (13) for the annealed approximation is

$$f = \alpha G_A - s(\{R_{ij}, Q_{kl}\}) \quad (17)$$

where the rescaled number of examples $\alpha = P/N$ is independent of β. Unlike the high temperature limit, in the annealed approximation the empirical average $\frac{1}{P}\sum_{\mu=1}^{P} e(\mathbf{W}, \boldsymbol{\xi}^\mu)$ training set $I\!D$ is distinguished from $\langle e \rangle_\xi$. The training set performance is given by

$$e_{\text{train}} = \alpha^{-1} \frac{\partial}{\partial \beta} f(\{R_{ij}, Q_{kl}\}) \quad (18)$$

which has to be evaluated in the minimum of f.

5. Two-prototype systems

Here we discuss typical properties of the considered model situations as computed in the statistical physics framework. We first concentrate on the system with only two prototypes, which already displays non-trivial phenomena. Furthermore, significant differences between WTA and NG training can be observed.

Most of the discussion will be in terms of the high-temperature limit. We show, however, that the extension to lower temperatures by means of the Annealed Approximation gives similar results, qualitatively.

5.1. *Relevant configurations and minima of f*

In a system with two prototypes, successful learning should lead to the representation of each cluster by one of the $\boldsymbol{w}_i$. However, our analysis of WTA

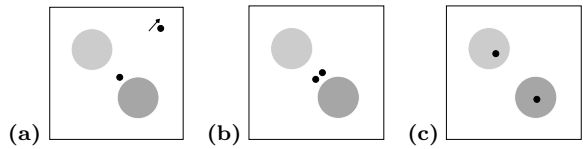


Fig. 1. Relevant configurations in the WTA scenario with two prototypes. The darker circle symbolizes the Gaussian cluster with larger prior weight. Graph (a) displays a configuration with one of the prototypes diverging. An unspecialized two-prototype configuration is shown in (b) and the optimal, specialized state is shown in panel (c). Note that the two prototypes in (b) would indeed coincide in the projection but are separated in the $(N-2)$-dim. orthogonal space.

training shows that this type of configuration competes with two other settings in thermal equilibrium. Figure 1 displays a sketch of all relevant situations: In configurations of type (a) only one of the prototypes is placed near the clusters while the second one diverges to infinity. Case (b) displays a situation with both prototypes in the region of high density but *unspecialized*, i.e. the specialization factor

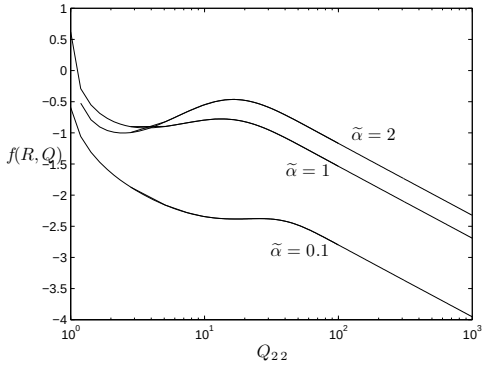
$$\Delta_m \;=\; |R_{1m} - R_{2m}| \tag{19}$$

is zero for all m. Panel (c) represents prototype configurations with $\Delta_m > 0$ which we will refer to as *specialized*.

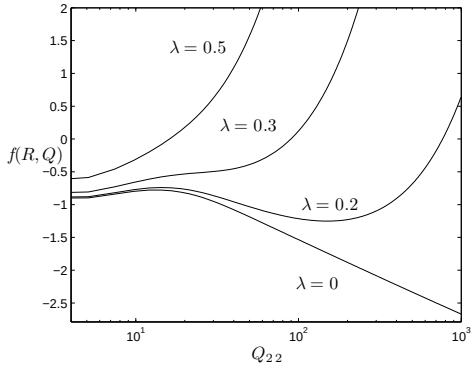
We first investigate the role of the trivial situation, (a), in the thermal equilibrium of WTA systems. In Fig. 2 (a) we have used the squared length Q_{22} of the second prototype as a parameter, while the free energy function is minimized with respect to all other order parameters. For all $\widetilde{\alpha}$, the global minimum of f, i.e. the true equilibrium state, corresponds to a trivial configuration: Only $\mathbf{w}_1$ contributes significantly to the energy $\widetilde{\alpha}\langle e\rangle_\xi$, while $\mathbf{w}_2$ does not represent any data as $Q_{22} \to \infty$. Note that the entropy s grows like $\ln Q_{22}$ while the energy approaches a constant value in this limit. Consequently, a trivial minimum of type (a) in Fig. 1 with $f \to -\infty$ will always be present.

For large enough data sets, a local minimum appears at smaller Q_{22} where both prototypes play a non-trivial role, e.g. $\widetilde{\alpha} = 1, 2$ in Fig. 2 (a). This corresponds to illustration (b) in Fig. 1. Both prototypes $\boldsymbol{w}_{1,2}$ coincide in the space spanned by $\boldsymbol{B}_{1,2}$, differences in the $(N-2)$-dim. orthogonal space are reflected by non-trivial configurations of $\{Q_{ij}\}$.

While the trivial configuration remains the true equilibrium state, local minima of the free energy are also relevant from a practical point of view: in a dynamical system approaching equilibrium, they correspond to metastable states. The time to leave



(a)



(b)

Fig. 2. Free energy function f vs. Q_{22} in a two-prototype system at $\beta \to 0$. The cost functions are (a) WTA for $\widetilde{\alpha} = 2$, 1 and 0.1, (b) NG with $\lambda = 0.5$, 0.3, 0.2 and 0 at $\widetilde{\alpha} = 1$. For both cases, $p_1 = p_2 = 0.5$, and $\ell = 1$.

a metastable state increases with the height of the free energy barriers confining it. The system can be *trapped* in a metastable configuration and typical *escape times* become prohibitively long in large systems, see e.g. [17] for a discussion.

This finding corresponds to the observation that initialization is highly important in practical applications of WTA-based systems. Given a start configuration, the system will approach and reside in the nearest stable or metastable state. Consequently, we expect the local minimum to be relevant in situations where the prototypes are prepared close to the clusters. On the contrary, prototypes initialized in regions with very low density of data will receive virtually no updates and do not contribute to the representation of data.

In rank based updates, as for instance in Neural Gas algorithms, the situation should be more favorable with respect to initialization issues. Here, all

prototypes are updated even if they are far away from the presented data.

The corresponding high temperature analysis of NG training shows that the trivial minimum with $Q_{22} \to \infty$ disappears for all $\lambda > 0$, as shown in Fig.2(b). The limit $\lambda \to 0$ is identical to WTA. In NG, all prototypes contribute to the extensive energy by a term on the order $g_\lambda(r_k)Q_{kk}$ which grows faster than entropy at large Q_{kk}. The trivial minimum is replaced by a local or global minimum at large but finite Q_{22}. The latter disappears completely at large enough values of the control parameter λ. The corresponding characteristic values of λ depend on the training set size $\widetilde{\alpha}$, the cluster geometry and parameters of the model density. Hence, NG systems can be expected to be less sensitive to initialization in practice. For large enough λ, the system is forced to place prototypes close to the cluster centers, cf. Fig.2(b). This reflects the benefits of annealing schemes in practical training with, initially, large λ to ensure that all prototypes converge.

In practice, the trivial states could also be avoided by means of setting proper boundaries to Q_{kk} or imposing a normalization. The latter would correspond to methods of directional clustering.

5.2. *Specialization transition in the training process*

The model parameters and the size of the training set determine which of the above discussed configurations are observed.

We first investigate in greater detail the WTA cost function with $\lambda = 0$. For small $\widetilde{\alpha}$, only trivial configurations, (a) in Fig. 1, are stable. Above a characteristic value of $\widetilde{\alpha}$, a non-trivial unspecialized state of type (b) becomes metastable. We assume now that the system resides in such a configuration and that the divergence of the second prototype has been avoided.

Fig. 3 shows that for small values of $\widetilde{\alpha}$ the prototypes remain unspecialized, i.e. $\Delta_m = 0$ for all m, see Eq. (19). 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 [10,16]. 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. 3 (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. 3 (b). The critical value depends on the model settings. For instance, $\widetilde{\alpha}_c$ will be larger for smaller ℓ.

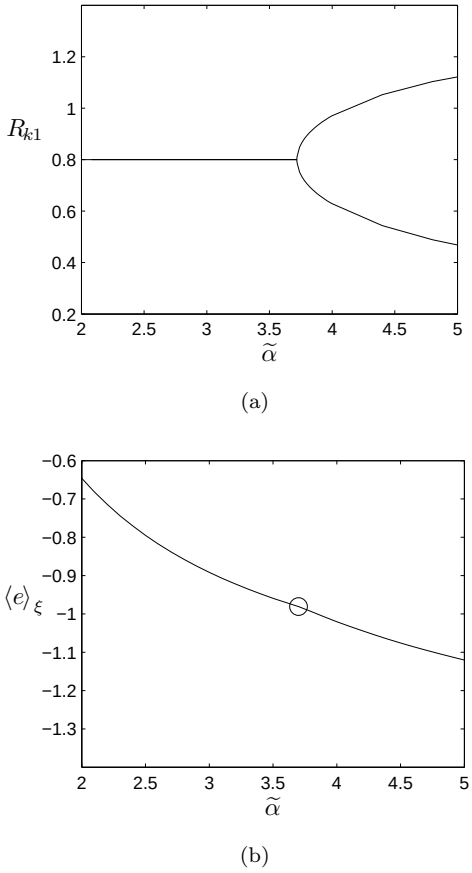


Fig. 3. (a) The order parameters R_{k1} of the stable configuration given the number of example $\widetilde{\alpha}$ for $K=2$. The system undergoes a continuous phase transition at a critical value $\widetilde{\alpha}_c(K=2) \approx 3.70$. The parameters of the input density (5) are $p_1 = 0.8, p_2 = 0.2$ and $\ell = 1$ in both panels. (b) The corresponding mean error $\langle e_{VQ} \rangle_\xi$ for $K=2$. The transition results in a kink for $K=2$ at $\widetilde{\alpha}_c$.

In summary, the generic behavior of the two-prototype WTA system is characterized by a se-

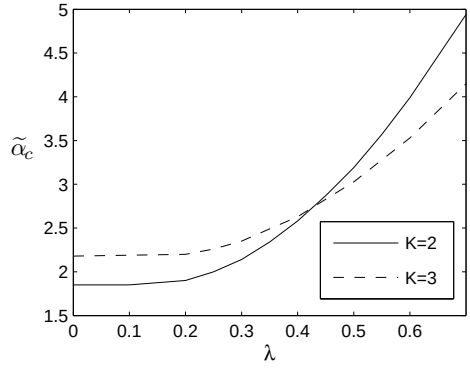


Fig. 4. Critical training set size α_c as a function of the rank control parameter λ, cf. Eq. (3). The parameters are set at $\ell = 1$, $p_1 = 0.8$, $p_2 = 0.2$.

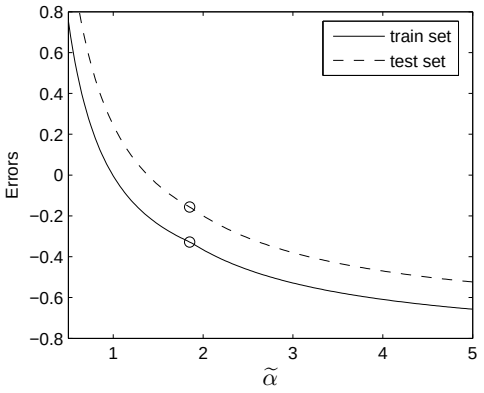
quence (a) → (b) → (c) with growing $\widetilde{\alpha}$ in terms of the illustrations in Fig. 1.

Next we investigate the minimization of rank based cost functions with $\lambda > 0$ in Eq. (1). We observe the same behavior as in WTA learning at sufficiently small rank function parameter λ. However, the critical value $\widetilde{\alpha}_c$ needed for prototype specialization and thus successful training, increases with λ, see Fig. 4. On the other hand, as discussed in Sec. 5.1, choosing large values of λ has the advantage of avoiding the divergence of one of the prototypes. 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.

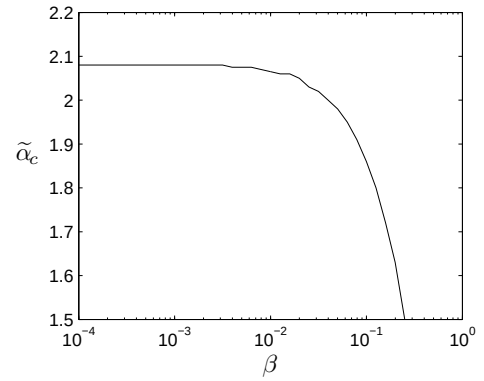
5.3. *Annealed approximation*

We have investigated the specialization transition in the WTA system with $K = 2$ using the annealed approximation. This approach allows us to study the learning behavior at finite temperatures. In the annealed approximation, inverse temperature β and training set size α can be chosen independently and the performance with respect to training and test data are distinguished. Figure 5(a) shows the average training errors and test errors for $\beta = 0.1$. The difference between e_{train} and $\langle e \rangle_\xi$ decreases with decreasing β and coincides with the results for the high temperature limit as $\beta \to 0$.

The annealed approximation exhibits qualitatively similar behavior as the high temperature limit: A continuous phase transition exists from unspecialized to specialized states. The continuous phase transition results in a kink in the learning



(a)



(b)

Fig. 5. Results for a WTA-based system with $K = 2$ using the annealed approximation. (a) The average errors for training set and test set as function of $\widetilde{\alpha}$ for $\beta = 0.1$. (b) The scaled critical training set sizes $\widetilde{\alpha}$ as function of β. The parameters of the input density are $p_1 = p_2 = 0.5$, $\ell = 1$.

curve for both the training error and test error.

For small β, we observe that $\alpha_c \propto \beta^{-1}$ which confirms the scaling $\widetilde{\alpha} = \beta\,\alpha$ in the limit $\beta \to 0$. Furthermore, figure 5(b) shows $\widetilde{\alpha}_c$ for a wider range of temperature. The annealed approximation predicts that the specialization transition is still relevant at finite temperatures, with $\widetilde{\alpha}_c$ decreasing with increasing β.

While the annealed approximation becomes exact in the limit $\beta \to 0$, it cannot be expected to describe the low temperature regime accurately. However, it has been frequently confirmed to reproduce qualitatively correct behavior [7,15]. The exact calculation of quenched averages at low temperatures would require, for instance, the the full replica approach. Here we conclude that the simplifying high temperature limit already gives reliable insight into

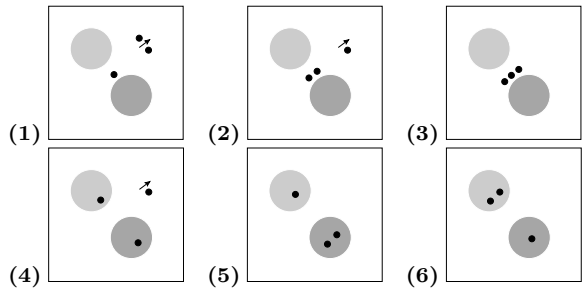


Fig. 6. Possible configurations in the WTA scenario with three prototypes. The darker circle symbolizes the Gaussian cluster with larger prior weight. Graphs(1-3) displays unspecialized configurations, with two prototypes diverging (1), one prototype diverging (2) and no prototypes diverging (3). Specialized configurations are shown in graphs (4-6), with one prototype diverging (4), low quantization error (5) and suboptimal performance (6). Note that unspecialized prototypes or prototypes representing the same cluster would indeed coincide in the projection but separate in the $(N-2)$-dim. orthogonal space.

the qualitative behavior of the model and proceed by applying it to more complex three prototype systems.

6. Three-prototype systems

The behavior of the phase transitions for $K = 3$ is qualitatively different compared to systems with $K = 2$. Figures 6 (1-6) show the relevant configurations which represent local or global minima of f. Depending on the cluster structure, for instance, the separation ℓ, distinct transition scenarios are observed.

6.1. *Small cluster separation*

The first example we analyze is a WTA system with $p_1 = 0.8, p_2 = 0.2$ and $\ell = 1$.

Similar to the two-prototype WTA system, for all $\widetilde{\alpha}$, the trivial global minimum of f is the configuration with only one prototype converging at the region of high density and two others diverging (Fig. 6(1)). At their respective characteristic $\widetilde{\alpha}$, new minima appear, in order, with two converging prototypes (Fig 6(2)) and three converging prototypes (Fig 6(3)). The relevant metastable configurations so far remain unspecialized.

At a critical value $\widetilde{\alpha}_s$, a specialized configuration with lower $\langle e_{VQ}\rangle_\xi$ appears. However, the phase transition is discontinuous or *first order*, i.e. a sudden jump occurs from an unspecialized to a specialized state, see Fig. 7. This behavior was also observed in

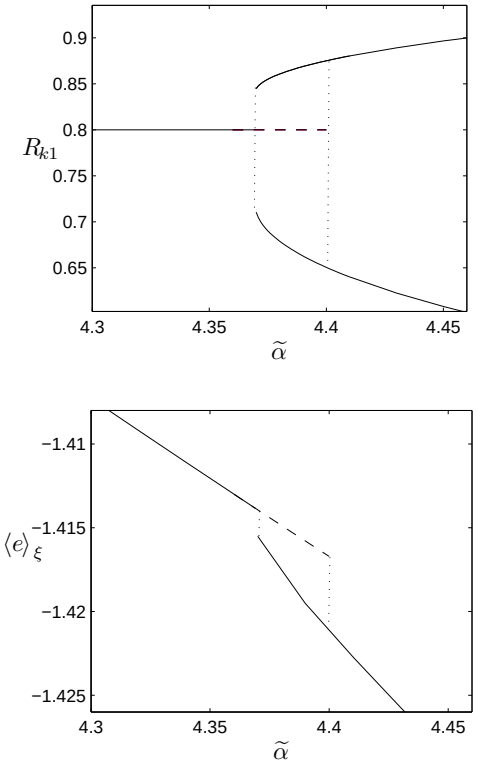


Fig. 7. (a) The order parameters R_{k1} of the stable configuration for $K = 3$, with two of the three values coinciding in the upper curve. The transition is discontinuous; solid (dashed) lines mark global (local) minima of f. Here, $\widetilde{\alpha}_s(K=3) \approx \widetilde{\alpha}_c(K=3) \approx 4.37$ and $\widetilde{\alpha}_d(K=3) \approx 4.40$. (b) The mean error $\langle e\rangle_{\boldsymbol{\xi}}$. The transition results in a discontinuous drop for $K=3$ at the respective $\widetilde{\alpha}_c$. The parameters of the input density (5) are $p_1=0.8, p_2=0.2$ and $\ell=1$ in both panels.

multilayer neural networks with three or more hidden units [5]. 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. Thus the behavior in this example is as follows, illustrated in Figs. 6: (1) → (2) → (3) → (5).

Note that, in a generic discontinuous phase transition, one expects a range of values $\widetilde{\alpha}_s \le \widetilde{\alpha} < \widetilde{\alpha}_c$ where the specialized configuration corresponds to a local minimum of f, see [5] for an example in supervised learning. 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 $\Delta_k = 0$. However, a local minimum of f corresponding to unspecialized $\mathbf{w}_i$ persists in the range $\widetilde{\alpha}_c \le \widetilde{\alpha} < \widetilde{\alpha}_d$. While such a

metastable state does not represent thermal equilibrium, its existence can have strong delaying effects in the practical optimization of H.

6.2. *Large cluster separation*

A different behavior can be observed in the $K = 3$ system in a scenario with relatively large separation, e.g. $\ell = 4$. Again at small $\widetilde{\alpha}$, metastable configurations (1) and (2) in Fig. 6 appear as $\widetilde{\alpha}$ grows.

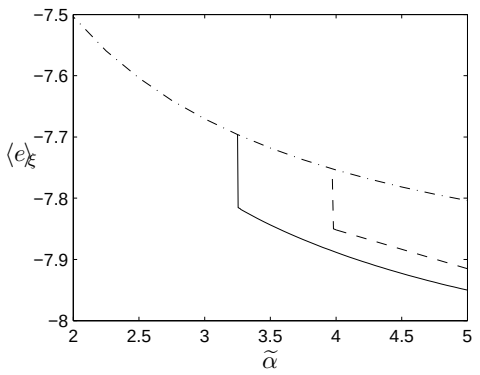
However, the specialization transition occurs at a much earlier stage, as the clusters are more easily identified. A specialized two-prototype configuration (Fig. 6(4)) then becomes metastable, even though no non-trivial three-prototype configuration exists. The latter emerges directly in a specialized state (Fig. 6(5)) at a lower quantization error. Here the three-prototype unspecialized configuration state is absent: the sequence is (1) → (2) → (4) → (5) as illustrated in Figs. 6.

An important feature in this scenario is the presence of local minima of $H(\mathbf{W})$. The lowest non-trivial minimum of f, i.e. the optimal configuration of prototypes, has two prototypes representing the cluster with larger prior, see Fig. 6(5) for an illustration. A suboptimal local minimum of f corresponds to the inverted configuration with two prototypes representing the weaker cluster, see Fig. 6(6). The sequence (1) → (2) → (4) → (6) in Figs. 6 can also be observed.

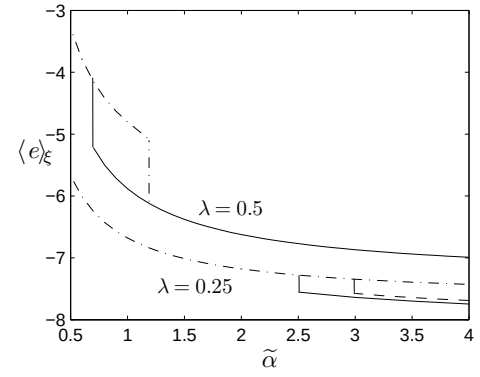
In this setting, the minima with configurations of type (5) and (6) are similar to global and local minima of the energy $H(\mathbf{W})$, respectively. Due to its lower value of f, the optimal state is also the more stable state at all $\widetilde{\alpha}$. However, in a dynamical context, the system may still be trapped near the local minima of $H(\mathbf{W})$ for long learning times.

6.3. *Neural Gas*

Figure 8 (b) relates to an NG system with $\lambda = 0.25$ and $\lambda = 0.5$ using the previously described example with larger separated clusters, $\ell = 4$. The characteristics of the system with $\lambda = 0.25$ are similar to WTA: non-trivial optimal and suboptimal configurations appear at certain values of $\widetilde{\alpha}$, depending on λ. Note that while NG with large λ forces the prototypes to be in a non-trivial state, it may also has its drawback in terms of $\langle e \rangle_{\boldsymbol{\xi}}$, i.e. it is higher than with WTA.



(a)



(b)

Fig. 8. Quantization error as function of $\widetilde{\alpha}$ for a $K = 3$ system with (a) WTA and (b) NG with $\lambda = 0.25$ and $\lambda = 0.5$. Here $\ell = 4$, $p_1 = 0.55$, $p_2 = 0.45$. The chained, solid and dashed lines correspond to the illustrations in Figs. 6 (4,5,6), respectively.

For sufficiently large λ, the energy landscape of $H(\mathbf{W})$ changes drastically. In contrast to WTA and NG with small λ, the suboptimal local minimum of f disappears, see Figure 8(b) for $\lambda = 0.5$. Thus, the parameter λ affects the number of local minima present in f and the smoothness of the energy landscape. The absence of the corresponding metastable states suggests that the NG-based system is robust, i.e. insensitive with respect to initial conditions in practice. This gives NG an advantage over WTA schemes, analogous to findings for on-line NG algorithms [20].

7. Summary

We have investigated the equilibrium properties of WTA and rank-based VQ systems along the lines

of statistical physics of off-line learning. The analysis of the learning behavior is based on two approaches: the high temperature limit and the annealed approximation. The simplifying high temperature limit provides exact analyses, yet already demonstrates the qualitative behavior of the annealed approximation at finite temperatures.

Both methods allow the study of the landscape of the cost function, which provides important insights into the training process. While the analysis concerns equilibrium properties of a hypothetical training process, its results are indeed relevant for practical situations. The existence of local metastable states, for instance, can influence their outcome drastically.

For, both, two- and three-prototype systems, a critical number of examples is required before the underlying structure can be detected at all. While it is obvious that any practical algorithm should give better performance with larger data sets, our results imply a more drastic effect: even the best optimization strategies will fail below this critical number of data. 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 metastable states for $K \geq 3$ show that long delays may happen in practice even if the critical number of examples is exceeded.

In WTA-based systems, divergent behavior may be observed if the training set is too small or the prototypes are initialized far from the region of high density. Furthermore, the system can be trapped in suboptimal local minima of the cost function. The NG-based system is more robust, i.e. for practical algorithms one can expect NG to be less sensitive to initial conditions. The outcome of NG training with regard to finding the optimal configuration will be studied in greater detail in upcoming papers.

Note that the phase transitions discussed here are found in the high temperature limit and annealed approximation, which makes qualitatively correct predictions for high temperatures. In the exact analysis of training at low temperatures, the nature of the phase transitions may be different. This could be treated in further extensions using the replica approach.

The formalism explained in this paper can also be applied to supervised learning schemes based on cost functions, which will be the focus of forthcoming projects.

Appendix A. Calculation of $\langle \ln Z \rangle_{I\!D}$ in the high-T limit

The partition sum defined in (6) is

$$Z = \int d\mu(\mathbf{W}) \, \exp\left[-\beta \, H(\mathbf{W})\right].$$

The computation of $\langle \ln Z \rangle_{I\!D}$ requires involved techniques from the theory of disordered systems. A common technique to perform the average over $I\!D$ is the so-called replica method, which exploits the relation

$$\langle \ln Z \rangle_{I\!D} = \lim_{n \to 0} \frac{1}{n} \ln \langle Z^n \rangle_{I\!D} \, . \tag{A.1}$$

For integer n, Z^n corresponds to the partition sum of n non-interacting replicas, i.e. identical copies of the system labeled $\gamma = 1, 2, \ldots, n$. From (6), we can rewrite

$$\langle Z^n \rangle_{I\!D} = \int \left[\prod_{\gamma} d\mu(\mathbf{W}^\gamma) \right] \, \exp\left[-\beta \sum_{\gamma=1}^{n} H\left(\mathbf{W}^\gamma\right) \right]. \tag{A.2}$$

Replacing $H(\mathbf{W}^\gamma)$ by an *effective* energy in the replicated space, (A.2) becomes

$$\langle Z^n \rangle_{I\!D} = \int \left[\prod_{\gamma} d\mu(\mathbf{W}^\gamma) \right] \exp\left[-\beta \, H_{\text{eff}} \left(\{\mathbf{W}^\gamma\}_{\gamma=1}^{n} \right) \right], \tag{A.3}$$

where H_{eff} is given by

$$H_{\text{eff}} \left(\{\mathbf{W}^\gamma\}_{\gamma=1}^{n} \right) = -\frac{1}{\beta} \ln \left\langle \exp\left[-\beta \sum_{\gamma=1}^{n} H\left(\mathbf{W}^\gamma\right) \right] \right\rangle_{I\!D} . \tag{A.4}$$

Because the examples are independent, we can treat $\langle . \rangle_{I\!D}$ separately for $\mu = 1, 2, \ldots, P$. Substituting $\langle . \rangle_{I\!D} = \langle . \rangle_{\xi}^{P}$, we obtain (A.4) as an average over the density described in (5),

$$H_{\text{eff}} \left(\{\mathbf{W}^\gamma\}_{\gamma=1}^{n} \right) = -\frac{P}{\beta} \ln \left\langle \exp\left[-\beta \sum_{\gamma=1}^{n} e(\mathbf{W}^\gamma, \boldsymbol{\xi}^\mu) \right] \right\rangle_{\xi} . \tag{A.5}$$

Note that different replicas interact in H_{eff} and are generally hard to be computed.

Here we resort to the high temperature limit $\beta \to 0$ which allows two simplifying relations [7,13,17]. First, the replicas become uncoupled, see e.g. [13]. The n-fold integral over $\prod_\gamma d\mu(\mathbf{W}^\gamma)$ are simply n multiplications between identical one-fold integral over $d\mu(\mathbf{W})$. Furthermore, $\langle \exp[-\beta H(\mathbf{W})] \rangle_{\mathcal{D}} = \exp\left[-\beta \left\langle H(\mathbf{W}) \right\rangle_{\mathcal{D}}\right]$. Equation (A.3) becomes greatly simplified as

$$\left\langle Z^n \right\rangle_{\mathcal{D}} = \left(\int d\mu(\mathbf{W}) \exp\left[-\beta P \left\langle e(\mathbf{W}, \boldsymbol{\xi}^\mu) \right\rangle_\xi \right] \right)^n \quad \text{(A.6)}$$

see [18] for more detailed calculations. Using the rescale $\widetilde{\alpha} = \beta(P/N)$ as in (9), and inserting (A.6) into the replica relations (A.1), we obtain the form

$$\left\langle \ln Z \right\rangle_{\mathcal{D}} = \ln \int d\mu(\mathbf{W}) \exp\left[-\widetilde{\alpha} N \left\langle e(\mathbf{W}, \boldsymbol{\xi}^\mu) \right\rangle_\xi \right]. \quad \text{(A.7)}$$

Next, we describe Eq. (A.7) in terms of order parameters in (11). By inserting the integral over the order parameters, it becomes

$$\begin{aligned} \left\langle \ln Z \right\rangle_{\mathcal{D}} = \ln \int d\mu(\mathbf{W}) & \int \left[\prod_{i,j} dR_{ij} \right] \left[\prod_{k,l \le k} dQ_{kl} \right] \\ & \times \prod_{i,j} \delta\left(R_{ij} - \mathbf{w}_i \cdot \mathbf{B}_j \right) \prod_{k,l \le k} \delta\left(Q_{kl} - \mathbf{w}_k \cdot \mathbf{w}_l \right) \\ & \times \exp\left[-\widetilde{\alpha} N \left\langle e \right\rangle_{\boldsymbol{\xi}} \right] \end{aligned} \quad \text{(A.8)}$$

The description of $\langle e \rangle_\xi$ in terms of order parameters $\{R_{ij}, Q_{kl}\}$ can be found in [3,4] for two-prototype systems, and [20] for three-prototype systems. We introduce the entropy term s which represents the phase space volume corresponding to a particular setting of order parameters.

$$\begin{aligned} s\left(\{R_{ij}, Q_{kl}\}\right) &= \frac{1}{N} \ln \int d\mu(\mathbf{W}) \prod_{i,j} \delta\left(R_{ij} - \mathbf{w}_i \cdot \mathbf{B}_j \right) \\ &\quad \times \prod_{k,l \le k} \delta\left(Q_{kl} - \mathbf{w}_k \cdot \mathbf{w}_l \right) \\ &= \frac{1}{2} \ln \det \mathsf{C} + c \end{aligned} \quad \text{(A.9)}$$

where c is independent of the order parameters and C is the covariance matrix

$$\mathsf{C} = \begin{pmatrix} Q_{11} & \cdots & Q_{1K} & R_{11} & R_{12} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ Q_{1K} & \cdots & Q_{KK} & R_{K1} & R_{K2} \\ R_{11} & \cdots & R_{K1} & 1 & 0 \\ R_{12} & \cdots & R_{K2} & 0 & 1 \end{pmatrix}, \quad \text{(A.10)}$$

see [1,18] for details. We substitute (A.9) into (A.8), to obtain the final form

$$\begin{aligned} \left\langle \ln Z \right\rangle_{\mathcal{D}} = \ln \int & \left[\prod_{i,j} dR_{ij} \right] \left[\prod_{k,l \le k} dQ_{kl} \right] \\ & \times \exp\left(-N \left[\widetilde{\alpha} \left\langle e \right\rangle_\xi - s\left(\{R_{ij}, Q_{kl}\}\right) \right] \right). \end{aligned} \quad \text{(A.11)}$$

References

[1] M. Ahr, M. Biehl, R. Urbanczik, Statistical physics and practical training of soft-committee machines, The European Physical Journal B 10 (1999) 583–588.

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

[3] M. Biehl, A. Freking, A. Ghosh, G. Reents, A theoretical framework for analysing the dynamics of LVQ: A statistical physics approach, Technical Report 2004-9-02, Mathematics and Computing Science, University Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands, December 2004, available from http://www.cs.rug.nl/~biehl.

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

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

[6] A. Buhot, M. B. Gordon, J.-P. Nadal, Rigorous bounds to retarded learning, Phys. Rev. Lett. 88 (9) (2002) 099801.

[7] A. Engel, C. van den Broeck, The Statistical Mechanics of Learning, Cambridge University Press, Cambridge, UK, 2001.

[8] D. Herschkowitz, M. Opper, Retarded learning: Rigorous results from statistical mechanics, Phys. Rev. Lett. 86 (10) (2001) 2174–2177.

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

[10] E. Lootens, C. van den Broeck, Analysing cluster formation by replica method, Europhys. Lett. 30 (1995) 381–387.

[11] T. Martinetz, S. Berkovich, K. Schulten, 'neural gas' network for vector quantization and its application to time series prediction, IEEE TNN 4(4) (1993) 558–569.

[12] Neural Networks Research Centre, Helsinki, Bibliography on the self-organizing

maps (SOM) and learning vector quantization (LVQ), Otaniemi: Helsinki Univ. of Technology. Available online: http://liinwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html .

[13] H. Seung, H. Sompolinsky, N. Tishby, Statistical mechanics of learning from examples, Physical Review A 45 (8) (1992) 6056–6091.

[14] S. A. Solla, E. Levin, Learning in linear neural networks: The validity of the annealed approximation, Physical Review A 46 (1992) 2124–2130.

[15] H. Sompolinsky, N. Tishby, Learning in a two-layer neural network of edge detectors, Europhys. Lett. 13:6 (1990) 567–572.

[16] T. Watkin, J. Nadal, Optimal unsupervised learning, J. Phys. A 27 (1994) 1899–1915.

[17] T. Watkin, A. Rau, M. Biehl, The statistical mechanics of learning a rule, Reviews of Modern Physics 65 (2) (1993) 499–556.

[18] 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 (2008).

[19] A. Witoelar, M. Biehl, A. Ghosh, B. Hammer, On the dynamics of vector quantization and neural gas, in: M. Verleysen (ed.), European Symposium on Artificial Neural Networks (ESANN) 2007, d-side, Evere, Belgium, 2007.

[20] A. Witoelar, M. Biehl, A. Ghosh, B. Hammer, Learning dynamics and robustness of vector quantization and neural gas, Neurocomputing 71 (2008) 1210–1219.

Aree Witoelar is currently a Ph.D. candidate in the Intelligent Systems Group of the Institute of Mathematics and Computing Science in University of Groningen, the Netherlands. He received his B.S. Degree in Engineering Physics from Bandung Institute of Technology, Indonesia, in 2002 and his M.Sc. Degree in Physics from University of Groningen, the Netherlands in 2005. His research interest is in the theory of machine learning, pattern recognition, clustering and self-organising maps.

Michael Biehl received his Ph.D. in Theoretical Physics from the University of Giessen, Germany, in 1992 and the venia legendi in Theoretical Physics from the University of Wurzburg, Germany, in 1996. In 2003 he was appointed Assistant Professor in Computing Science at the University of Groningen, The Netherlands. His main research interest is in the theory and application of Machine Learning techniques. He is furthermore active in the modelling and simulation of complex physical systems. He has co-authored more than 90 papers in international journals and conference proceedings; preprint versions and further information are available at http://www.cs.rug.nl/˜biehl