ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Equilibrium properties of offline LVQ

Aree Witoelar[1], Michael Biehl[1], 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. The statistical physics analysis of offline learning is applied to cost function based learning vector quantization (LVQ) schemes. Typical learning behavior is obtained from a model with data drawn from high dimensional Gaussian mixtures and a system of two or three competing prototypes. The analytic approach becomes exact in the limit of high training temperature. We study two cost function related LVQ algorithms and the influence of an appropriate weight decay. In our findings, learning from mistakes (LFM) achieves poor generalization ability, while a limiting case of generalized LVQ (GLVQ), termed LVQ+/-, displays much better performance with a properly chosen weight decay.

1 Introduction

Learning Vector Quantization (LVQ) and other prototype based classification methods have been successfully applied in various fields, e.g. data mining, medical image analysis and speech recognition, see [6] for an extensive bibliography.

In many LVQ schemes, the goal of good generalization behavior is aimed at by minimizing a cost function which is expected to yield low error rates. Such cost functions can be inspired by the concept of large margins as in, e.g., Generalized LVQ (GLVQ) [7] or relate to explicit assumptions about the statistics of data, for instance Robust Soft LVQ (RSLVQ) [8]. While these cost functions may appear plausible, their relation with the generalization ability remains unclear.

Methods from statistical physics allow to investigate equilibrium properties of large systems, such as neural networks [9, 10]. In the by now standard analysis of off-line (batch) learning, training is interpreted as stochastic minimization of a cost function. Here, our analysis is based on the limit of high training temperature, a simplification that has given insights into many learning scenarios [3, 9, 10, 11].

In this paper, we analyze two important LVQ schemes: a limiting case of GLVQ and Learning From Mistakes (LFM) in systems of two and three competing prototypes. Both learning schemes have been studied as on-line learning in [2], which are based on a sequence of single example data. Here we explicitly treat off-line learning from a given fixed data set. In addition we study *weight decay* as a control parameter against instabilities. Our analysis shows how successful learning depends on the size of the training set. In particular, the learning process exhibits phase transitions which influence the learning process, reminiscent of those observed in unsupervised VQ [11] or multilayer neural networks [1, 3, 9]. These findings provide useful insights into the behavior of general off-line VQ schemes.

2 Model Data

Consider a data set of P examples given as $I\!\!D = \{(\boldsymbol{\xi}^\mu, \sigma^\mu) \in I\!\!R^N \times \{+1,-1\}\}_{\mu=1}^P$. We exploit the thermodynamic limit $N \to \infty$ and assume that the number of examples also grows linearly in N, i.e. $P \propto N$. Examples are generated independently according to a given model density, a mixture of two Gaussians:

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

where the prior weights satisfy $p_+ + p_- = 1$. The cluster centers are given by $\ell\mathbf{B}_+$ and $\ell\mathbf{B}_-$, where ℓ is a separation parameter and $\mathbf{B}_m$ are orthonormal with $\mathbf{B}_m \cdot \mathbf{B}_n = \delta_{mn}$. Densities of the above form have been studied previously in the context of supervised and unsupervised learning, see e.g. [2, 4, 5, 12]. Note that the highly overlapping clusters separate only in the subspace spanned by the $\mathbf{B}_\pm$. Random two-dim. projections do not display any separation, see [2].

3 Cost functions

We consider a system of K prototype vectors $\mathbf{W} = \{(\mathbf{w}_k, c_k)\}_{k=1}^K$. Cost functions considered here are expressed as empirical averages $H(\mathbf{W}) = \sum_{\mu=1}^P e(\mathbf{W}, \boldsymbol{\xi}^\mu)/P$. We study the following specific examples:

- LVQ+/- with $e(\mathbf{W}, \boldsymbol{\xi}^\mu) = d_S^\mu - d_T^\mu$.
  We restrict the analysis to the Euclidean distance $d_k^\mu = (\mathbf{w}_k - \boldsymbol{\xi}^\mu)^2$. The prototype $\mathbf{w}_S$ is the closest correctly labeled prototype, while $\mathbf{w}_T$ is the closest incorrectly labeled prototype. This is similar to the LVQ 2.1 prescription which selects data from a *window* about the decision boundary. Here we study the limit of infinite window size. GLVQ [7] with $e(\mathbf{W}, \boldsymbol{\xi}^\mu) = \Phi\left[(d_S^\mu - d_T^\mu)/(d_S^\mu + d_T^\mu)\right]$ reduces to the above for $N \to \infty$ and $\Phi(x) = x$. Note that for $N \to \infty$ the term $\boldsymbol{\xi}^2 = \mathcal{O}(N)$ dominates $d_S^\mu - d_T^\mu$ yielding a constant prefactor, while it cancels out in the numerator $d_S^\mu - d_T^\mu$.

- LFM with $e(\mathbf{W}, \boldsymbol{\xi}^\mu) = (d_S^\mu - d_T^\mu)\,\Theta(d_S^\mu - d_T^\mu)$.
  The prototypes $\mathbf{w}_S$ and $\mathbf{w}_T$ are defined as in LVQ+/- and $\Theta(x)$ is the Heaviside function. Here only misclassified data contribute to the cost reminiscent of perceptron training [4, 10]. We refer to this as learning from mistakes (LFM).

4 Equilibrium Physics Approach

In the statistical physics analysis of off-line learning, [9, 10], training is interpreted as the stochastic minimization of $H(\mathbf{W})$ where a temperature parameter T controls the degree of randomness. This leads to a well-defined thermal equilibrium: a configuration $\mathbf{W}$ occurs with a probability given by the Gibbs density $P(\mathbf{W}) = \exp\left[-\beta H(\mathbf{W})\right]/Z$ where $Z = \int d\mu(\mathbf{W}) \exp\left[-\beta H(\mathbf{W})\right]$. 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 evaluated as derivatives of the so-called free energy $-\ln Z/\beta$.

Note that the above is defined for one specific data set. In order to obtain typical properties of the model scenario, an additional average over all $I\!\!D$ of the same size and statistical properties is performed. It yields the so-called quenched free energy $-\langle \ln Z \rangle_{I\!\!D} / \beta$ [4, 9, 10]. In general, the computation of $\langle \ln Z \rangle_{I\!\!D}$ requires involved techniques from the theory of disordered systems [9]. Here we resort to the simplifying limit of high temperatures, $\beta \to 0$, which has given valid insight into a variety of learning scenarios [4, 9, 10]. Non-trivial results can only be expected if the increased noise is compensated for by a large number of examples $P = \widetilde{\alpha} N/\beta$. Because large training sets sample the model density very well, $H(\mathbf{W})$ can be replaced by $P\ \langle e \rangle_{\xi}$, i.e. the average over $P(\boldsymbol{\xi})$. In the limit $N \to \infty$, $\langle e \rangle_{\xi}$ can be expressed as a function of very few order parameters

$$R_{ij} = \mathbf{w}_i \cdot \mathbf{B}_j \ \text{ and } \ Q_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j. \tag{2}$$

Note that R_{ij} are the projections of prototype vectors $\mathbf{w}_i$ on the center vectors $\mathbf{B}_j$ and Q_{ij} correspond to the self- and cross- overlaps of the prototype vectors. Here (2) defines $2K + K(K+1)/2$ order parameters for K prototypes, see [12] for the result and details of the calculation. It can be performed analytically for systems with two prototypes but involves numerical Gaussian integrals for $K \geq 3$. We rewrite $\langle \ln Z \rangle_{I\!\!D}$ as an integral over the order parameters:

$$\langle \ln Z \rangle_{I\!\!D} = \ln \int \prod_{i,j} dR_{ij} \prod_{i,j \leq i} dQ_{ij} \exp\left(-N \left[\widetilde{\alpha}\ \langle e \rangle_{\xi}\ -\ s\left(\{R_{ij}, Q_{ij}\}\right) \right] \right). \tag{3}$$

Here, the entropy s gives the phase space volume of a particular order parameter configuration [10]. In the limit $N \to \infty$ the free energy (3) is dominated by the maximum integrand, i.e. the minimum of $f(\{R_{ij}, Q_{ij}\}) = \widetilde{\alpha}\ \langle e \rangle_{\xi} - s\left(\{R_{ij}, Q_{ij}\}\right)$. Hence, given a specific cost function and training set size $\widetilde{\alpha}$, we obtain typical equilibrium properties by minimizing the free energy function $f(\{R_{ij}, Q_{ij}\})$ with respect to the order parameters. Also, the generalization error $\epsilon_g(\mathbf{W}) = \int d\boldsymbol{\xi} P(\boldsymbol{\xi}) \sum_{k:c^k \neq \sigma}^{K} \prod_{l \neq k}^{K} \Theta(d_l^{\mu} - d_k^{\mu})$, can be expressed in terms of $\{R_{ij}, Q_{ij}\}$ [12].

5 Results

We first present the result for a system with two prototypes for LFM. Minimizing the free energy function f, the equilibrium configuration is found at finite $\{R_{ij}(\widetilde{\alpha}), Q_{ij}(\widetilde{\alpha})\}$. Although LFM appears reasonable, it exhibits surprisingly poor generalization ability compared to the best linear decision boundary, displayed in Fig. 1. The non-monotonic behavior wrt. $\widetilde{\alpha}$ indicates that this cost function is not well suited to this particular training task. For highly unbalanced priors, ϵ_g even exceeds the trivial value $\min\{p_+, p_-\}$.

Meanwhile, the LVQ+/- system always exhibits divergent behavior. For unequal priors, only the trivial minimum $f = -\infty$ exists, with infinite length of the prototype representing the weaker class, i.e. $Q_{kk} \to \infty$. To avoid this known problem, we choose the regularization method by punishing configurations with large lengths using an additional energy called *weight decay* [1]. We place the origin of the weight decay at the center of mass, $\ell(p_+\mathbf{B}_+ + p_-\mathbf{B}_-)$, and preserve the symmetry axis. In practice, this is equivalent to transforming the data into zero

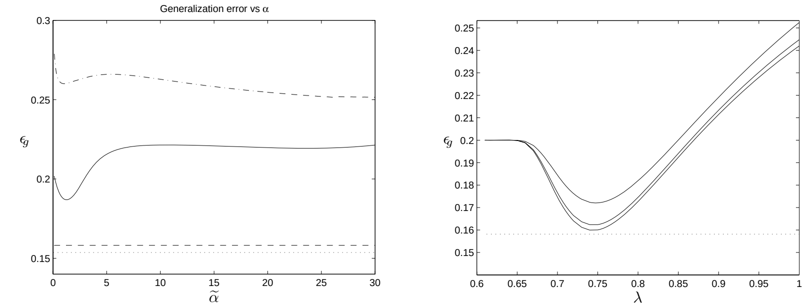

Fig. 1: Left panel: ϵ_g vs $\widetilde{\alpha}$ for LFM. The achieved error rate are displayed for $K = 2$ ($K = 3$) with solid (chain) lines. The dashed (dotted) lines mark the optimal errors using two (three) prototypes. Right panel: ϵ_g vs weight decay λ for LVQ+/- with $K = 2$ for $\widetilde{\alpha} = 1, 4$ and 10 (solid lines, top to bottom). The performance approaches the best linear decision boundary (dotted lines) as $\widetilde{\alpha} \to \infty$ with proper settings of λ. For both figures, $p_+ = 0.8$ and $\ell = 1$.

mean and calculating the weight decay wrt. the transformed origin. Hence we obtain the modified cost function $H = \sum_{\mu=1}^{P} e + \lambda P \sum_{k=1}^{K} [\mathbf{w}_k - \ell(p_+\mathbf{B}_+ + p_-\mathbf{B}_-)]^2$. The free energy function to be minimized becomes

$$f(\{R_{ij}, Q_{ij}\}) = \widetilde{\alpha} \left\langle e \right\rangle_\xi + \widetilde{\alpha}\lambda \sum_{k=1}^{K} \left(Q_{kk} - 2\ell\left(p_+ R_{k+} + p_- R_{k-}\right)\right) - s\left(\{R_{ij}, Q_{ij}\}\right). \quad (4)$$

The performance of LVQ+/- is improved with proper settings of λ, displayed in Fig. 1 (right). At settings with small λ, the prototypes have very large lengths, which is not the desired result of training. Conversely, large λ places greater importance of the weight decay at the expense of higher generalization error. The optimal λ is fairly robust wrt. the size of the training set.

The behavior described above differs from that of unsupervised vector quantization in [11]. In the latter, permutations between prototypes lead to effects of *retarded learning*, i.e. a minimum number of examples is required to have any chance of successful learning, see e.g. [3, 5]. Meanwhile in the supervised learning scenario here, the permutation symmetry between prototypes is broken by the class assignment of each prototype. Therefore, as long as $\widetilde{\alpha}$ is in the order $\mathcal{O}(1)$, each prototype already aligns itself towards its respective class mean.

The learning behavior is qualitatively different for systems with three prototypes. Possible permutations between two prototypes of the same class, e.g. $\mathbf{w}_k$ and $\mathbf{w}_l$, produce two distinct states, illustrated in Fig. 2 (left). In the first state, these prototypes have equal lengths, and lie symmetrically around $\ell(\mathbf{B}_+ - \mathbf{B}_-)$. All three prototypes form a wedge-shaped decision boundary which allows for better generalization ability, in general. For symmetrical reasons, we can represent this state with a configuration with $R_{km} = R_{lm}, \forall m$ and $Q_{kk} = Q_{ll}$. In the second state, the prototypes have unequal lengths and lie asymmetrically along $\ell(\mathbf{B}_+ - \mathbf{B}_-)$. The decision boundary is predominantly defined by only two prototypes. Here the prototype with the larger length can diverge, thus weight

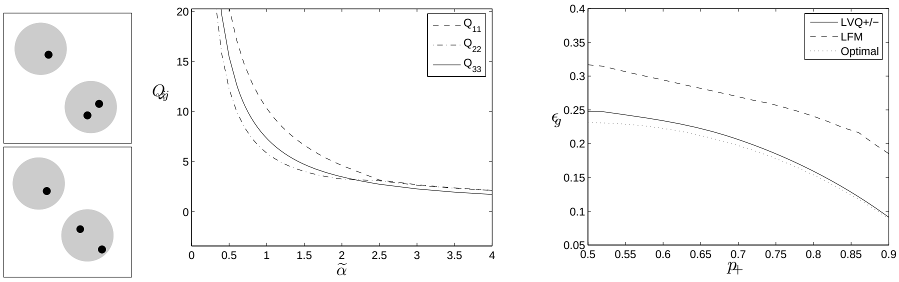

Fig. 2: Left panel: Two distinct configurations of a $K = 3$ system with two similarly labeled prototypes $c_k = \{+, +, -\}$. Middle panel: Vector lengths Q_{kk}, for LFM with $K = 3, p_+ = 0.8$ The system undergoes a continuous phase transition at $\widetilde{\alpha} \approx 2.5$. Right panel: ϵ_g vs. p_+ with $K = 3$, $\widetilde{\alpha} \to \infty$ and optimal settings of λ using LVQ+/- (solid lines) and LFM (dashed lines). The optimal error for three prototypes is displayed with dotted lines.

decay is necessary.

In the three-prototype LFM system with weight decay, a continuous phase transition occurs at a critical number of examples $\widetilde{\alpha}_c$, see Fig. 2 (middle). For settings of small λ and small $\widetilde{\alpha}$, we find the asymmetric configuration. At $\widetilde{\alpha}_c$, the system switches to the symmetric configuration. The configurations are illustrated by the second and first configuration in Fig. 2 (left), respectively. This translates into a non-differentiable kink in the learning curve. In general, the performance of LFM does not improve with weight decay. Also note that additional degrees of freedom and larger entropy of the three prototype system must be compensated for by using a larger number of examples to perform better than the two-prototype system.

Finally, we investigate LVQ+/- for three prototype systems. Due to larger repulsion from the stronger class, LVQ+/- requires larger λ than LFM to prevent divergence. With this decay, the asymmetric configuration is unfavorable in terms of free energy and therefore no such states are found. Consequently, in these model settings, the above mentioned phase transition is not observed in the learning process.

Note that while various performances of one specific cost function can be analyzed, comparison between two different cost functions is difficult in the high temperature limit. Any multipliers of $H(\mathbf{W})$ are rescaled into $\beta \to 0$ and, consequently, the scale of $\widetilde{\alpha}$ is not consistently defined between different cost functions except at the limit $\widetilde{\alpha} \to \infty$. In this limit, LVQ+/- outperforms LFM given a properly chosen weight decay, as shown in the right panel of Fig. 2.

6 Discussion

We have investigated LVQ systems for two important cost functions: LFM and LVQ+/-. The analysis is performed along the lines of the statistical physics treatment of off-line learning using the limit of high training temperatures. We expect the results from the analysis to carry over qualitatively to learning at low temperatures, similar to previous studies of supervised learning, e.g. [3, 9].

While LFM appears plausible, its performance is unexpectedly poor compared to the optimum achievable error for this learning problem. The achieved generalization error does not improve even with very large training sets. Meanwhile as expected, LVQ+/- displays divergent behavior and requires modifications such as a weight decay. Given properly chosen weight decay, LVQ+/- exhibits better generalization ability than LFM for both two- and three prototype systems.

In three prototype systems, we find continuous phase transitions between prototype configurations. While any practical algorithm should give better performance with larger data sets, a critical size of the training set is required to effectively utilize all available prototypes. Treatment of systems with more prototypes can lead to existence of other phase transitions, e.g. as found in [3, 11].

In future projects we will rigorously study and compare various cost function based LVQ algorithms including general forms of GLVQ and modifications such as window schemes. Also, non-trivial behavior of the system wrt. to weight decay settings was observed in this work and will be investigated. Finally, the analysis at finite temperatures allows independent variation of the number of examples P/N and T. This analysis is highly important for practical applications, e.g. simulated annealing schemes which ends with low temperatures.

References

[1] M. Ahr, M. Biehl, and E. Schloesser. Weight decay induced phase transitions in multilayer neural networks. *Journal of Physics A: Mathematical and General*, 32:5003–5008, 1999.

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

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

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

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

[6] Neural Networks Research Centre, Helsinki. Bibliography on the self-organizing maps (SOM) and learning vector quantization (LVQ). *Otaniemi: Helsinki Univ. of Technology. http://liinwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html* , 2002.

[7] A. Sato and K. Yamada. Generalized learning vector quantization. In *NIPS*, pages 423–429, 1995.

[8] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

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

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

[11] A. Witoelar and M. Biehl. Phase transitions in vector quantization and neural gas. *Neurocomputing*, 2009.

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