

## Abstract

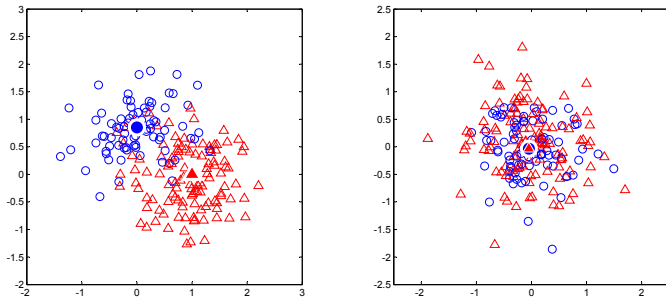
The statistical physics analysis of offline learning is applied to cost function based learning vector quantization (LVQ) schemes. Typical learning behavior is obtained from a model with data drawn from high dimensional Gaussian mixtures and a system of two or three prototypes. The analytic approach 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.

## Introduction

Learning Vector Quantization (LVQ) is an intuitive and powerful family of prototype-based learning algorithms, successfully applied in various fields. In many LVQ schemes, a cost function is minimized with expectations of low error rates. However, its relation with the generalization ability remains unclear. Methods from statistical physics allow to investigate large systems, such as neural networks. In the thermodynamic limit of high dimensions, the system can be described by a set of order parameters. Our analysis shows how successful learning depends on the size of the training set.

## Model Data and Cost Functions

Consider a data set of  $P$  examples given as  $D = \{\xi^\mu, \sigma^\mu\}_{\mu=1}^P$  where input vector  $\xi^\mu \in \mathbb{R}^N$  and class  $\sigma^\mu \in \pm 1$ . Examples are generated independently according to a given model density: a mixture of two spherical Gaussians with priors  $p_+, p_-$  and orthonormal cluster center vectors  $\mathbf{B}_+, \mathbf{B}_-$ .



**Figure 1:** Sample data generated by the model (here  $N = 100$ ). The clusters separate in the projection to the plane spanned by cluster center vectors  $\mathbf{B}_+, \mathbf{B}_-$  (Left) but completely overlap in the projection to an arbitrary plane (Right).

We consider a system of  $K$  prototype vectors  $\mathbf{W} = \{\mathbf{w}_k, c_k\}_{k=1}^K$  and cost function

$$H(\mathbf{W}) = \frac{1}{P} \sum_{\mu=1}^P e(\mathbf{W}, \xi^\mu) + \text{decay}$$

with the following specific examples:

- **LVQ+/-**:  $e(\mathbf{W}, \xi^\mu) = d_S^\mu - d_T^\mu$  with Euclidean distance  $d_k^\mu = (\mathbf{w}_k - \xi^\mu)^2$ . The prototypes  $\mathbf{w}_S$  and  $\mathbf{w}_T$  are the closest correct and incorrect labeled prototypes.
- **Learning From Mistakes (LFM)**:  $e(\mathbf{W}, \xi^\mu) = (d_S^\mu - d_T^\mu) \Theta(d_S^\mu - d_T^\mu)$  where  $\Theta(x)$  is the Heaviside function. Only misclassified data contribute to the cost.

In addition we study *weight decay* as a control parameter against instabilities by punishing  $\mathbf{W}$  with large lengths, i.e.  $\text{decay} \propto \lambda Q_{kk}$  where  $\lambda$  is a weight decay parameter.

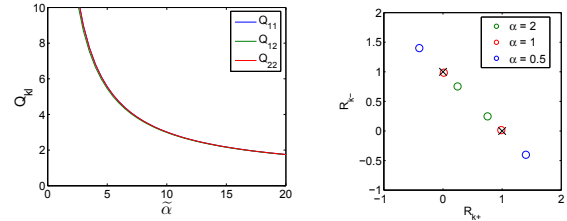
## Equilibrium Physics

The statistical physics analysis of off-line learning is outlined as follows:

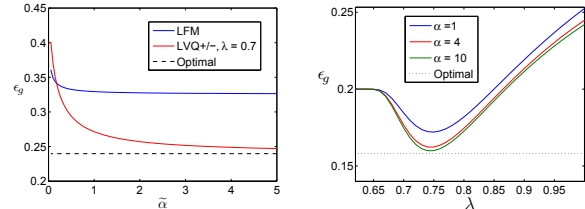
- training is interpreted as minimization of  $H(\mathbf{W})$  with temperature  $T$
- thermal equilibrium: configuration  $\mathbf{W}$  occurs with a probability  $P(\mathbf{W}) = \exp[-\beta H(\mathbf{W})]/Z$  where  $\beta = 1/T$  and the normalization  $Z$  is called the partition sum.
- typical learning properties can be obtained from  $(\ln Z)_D$ , which simplifies in the limit  $\beta \rightarrow 0$
- in the limit  $N \rightarrow \infty$ , the averages can be expressed as a function of very few order parameters  $R_{ij} = \mathbf{w}_i \cdot \mathbf{B}_j$  and  $Q_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j$
- given training set size  $P = \tilde{\alpha} N / \beta$ , certain  $\{R_{ij}, Q_{ij}\}$  configurations dominate the equilibrium density.

## Results

### Typical learning curves



**Figure 2:** Left: Order parameters  $Q_{kl}$  vs training set size  $\tilde{\alpha}$  for LFM,  $p_+ = 0.5$  ( $Q_{11}$  and  $Q_{22}$  coincide). Right:  $R_{k+}$  vs  $R_{k-}$  plots display the projection of prototypes on  $\text{span}(\mathbf{B}_+, \mathbf{B}_-)$  at training set sizes  $\tilde{\alpha} = 0.5, 1$  and  $2$ . Crosses mark the two class centers.

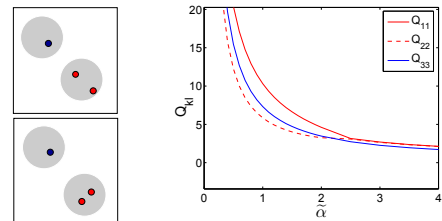


**Figure 3:** Left: Corresponding generalization error  $\epsilon_g$  vs  $\tilde{\alpha}$  of Fig. 2 for LFM and LVQ+/- with weight decay with  $K = 2$ . Right:  $\epsilon_g$  vs weight decay  $\lambda$  for LVQ+/- for  $\tilde{\alpha} = 1, 4$  and  $10$ ,  $p_+ = 0.8$ . The performance approaches the best linear decision boundary with proper settings of  $\lambda$  as  $\tilde{\alpha} \rightarrow \infty$ .

Given properly chosen weight decay, LVQ+/- exhibits better generalization ability than LFM for both two- and three-prototype systems.

### Phase transitions

For systems with more prototypes, possible permutations between prototypes produce distinct states which can dominate the training process.



**Figure 4:** Left: Two distinct configurations of a  $K = 3$  system with two similarly labeled prototypes  $c_k = \{+, +, -\}$ . Right: Vector lengths  $Q_{kk}$  for LFM with  $K = 3, p_+ = 0.8$ . The system undergoes a continuous phase transition at  $\tilde{\alpha} \approx 2.5$  from the top- to bottom configuration on the left panel.

## Outlook

In future projects we will study various cost function based LVQ algorithms. Also, we will extend the analysis to finite temperatures which provides important insights to practical applications. This analysis requires techniques such as the annealed approximation or replica method and allows independent variation of the number of examples  $P/N$  and temperature.

## References

- [1] Neural Networks Research Centre, Helsinki. Bibliography on the self-organizing maps (SOM) and learning vector quantization (LVQ). *Otaniemi: Helsinki Univ. of Technology*. <http://linwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html>, 2002.
- [2] A. Engel and C. van den Broeck. *The Statistical Mechanics of Learning*. Cambridge University Press, Cambridge, UK, 2001.
- [3] H.S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056--6091, 1992.
- [4] 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.
- [5] A. Witoelar and M. Biehl. Phase transitions in vector quantization and neural gas. *Neurocomputing*, 2009.