Dynamics and generalization ability of LVQ algorithms

Biehl, Michael; Ghosh, Anarta; Hammer, Barbara

Published in:
Journal of Machine Learning Research

Publication date:
2007

Journal of Machine Learning Research 8 (2007) 323-360 Submitted 9/05; Revised 7/06; Published 2/07

Dynamics and Generalization Ability of LVQ Algorithms

Michael Biehl M.BIEHL@RUG.NL
Anarta Ghosh ANARTA@GMAIL.COM
Institute for Mathematics and Computing Science
University of Groningen
P.O. Box 800, NL-9700 AV Groningen, The Netherlands

Barbara Hammer HAMMER@IN.TU-CLAUSTHAL.DE
Institute of Computer Science
Clausthal University of Technology
D-38678 Clausthal-Zellerfeld, Germany

Editor: Yoshua Bengio

Abstract

Learning vector quantization (LVQ) schemes constitute intuitive, powerful classification heuristics with numerous successful applications but, so far, limited theoretical background. We study LVQ rigorously within a simplifying model situation: two competing prototypes are trained from a sequence of examples drawn from a mixture of Gaussians. Concepts from statistical physics and the theory of on-line learning allow for an exact description of the training dynamics in high-dimensional feature space. The analysis yields typical learning curves, convergence properties, and achievable generalization abilities. This is also possible for heuristic training schemes which do not relate to a cost function. We compare the performance of several algorithms, including Kohonen's LVQ1 and LVQ+/-, a limiting case of LVQ2.1. The former shows close to optimal performance, while LVQ+/- displays divergent behavior. We investigate how early stopping can overcome this difficulty. Furthermore, we study a crisp version of robust soft LVQ, which was recently derived from a statistical formulation. Surprisingly, it exhibits relatively poor generalization. Performance improves if a window for the selection of data is introduced; the resulting algorithm corresponds to cost function based LVQ2. The dependence of these results on the model parameters, for example, prior class probabilities, is investigated systematically, simulations confirm our analytical findings.

Keywords: prototype based classification, learning vector quantization, Winner-Takes-All algorithms, on-line learning, competitive learning

1. Introduction

The term *learning vector quantization* (LVQ) has been coined for a family of algorithms which is widely used in the classification of potentially high-dimensional data. Successful applications of LVQ include such diverse problems like medical image or data analysis, fault detection in technical systems, or the classification of satellite spectral data (Bojer et al., 2003; Kuncheva, 2004; Schleif et al., 2006; Villmann et al., 2003). An up to date overview and extensive bibliography is available at a repository which is maintained by the Neural Networks Research Centre, Helsinki (2002).

The popularity of the approach is due to several attractive features: LVQ procedures are easy to implement and intuitively clear. This makes them particularly interesting for researchers and practitioners outside the machine learning community who are searching for robust classification schemes

without the black-box character of many neural methods. The classification of LVQ is based on a distance measure, usually the Euclidean distance, which quantifies the similarity of given data with so-called prototype or codebook vectors representing the classes. The prototypes are determined in a training process from labeled example data and can be interpreted in a straightforward way as they capture essential features of the data in the very same space. This is in contrast with, say, adaptive weights in feedforward neural networks or support vector machines which do not allow for immediate interpretation as easily since they are embedded in a different space or at atypical borderline positions of the data instead of typical regions. Other very attractive features of LVQ are the natural way in which it can be applied to multi-class problems and the fact that the complexity of LVQ networks can be controlled during training according to the specific needs.

In general, several prototypes can be employed to represent one of the classes. In simple *hard* or *crisp* schemes, a data or feature vector is assigned to the closest of all prototypes and the corresponding class. Extensions of this deterministic scheme to probabilistic *soft* assignments and classification are straightforward but will not be considered here. Plausible training prescriptions exist which mostly employ the concept of competitive learning and adaptation by means of Hebbian terms. Prototypes are updated according to their distance from given example data. Thereby, given a training pattern, the closest prototype vector or a set of closest vectors, the so-called *winners* are identified. These vectors are then moved towards (away from) the data if their class label coincides with (differs from) that of the example, respectively.

Both, training algorithms and the resulting classification schemes are fairly easy to implement and fast. In practice, the computational effort of LVQ training usually scales linearly with the training set size, and that of classification depends only on the (fixed) number of prototypes and the input dimensionality. Furthermore, training is incremental, such that the adaptation of a classifier to novel data which becomes available after a first training phase is straightforward. Despite this simplicity, LVQ is very powerful since every separating hypercurve can, in principle, be approximated by LVQ networks with sufficiently many prototypes. Furthermore, recent variations of LVQ allow for the incorporation of problem specific metrics or kernels which can be adapted during training such that very few prototypes can model quite complex behavior (Hammer et al., 2005b,c).

However, the theoretical understanding of the convergence properties, stability, and achievable generalization ability of many LVQ algorithms appears fairly limited. Many LVQ procedures, including Kohonen's original formulation (LVQ1) are based on heuristic arguments. A large variety of modifications of the basic scheme have been suggested which aim at larger flexibility, faster convergence or better approximation of Bayesian decision boundaries, such as LVQ2.1, LVQ3, OLVQ, RSLVQ, or GLVQ (Kohonen, 1997, 1990; Pregenzer et al., 1996; Sato and Yamada, 1995). Clearly, the ultimate goal of training is good generalization, that is, the correct classification of novel data with high probability after training. However, the above mentioned algorithms differ considerably in their learning dynamics, stability, and generalization performance, while the theoretical background of this behavior is hardly understood.

Recently, a few attempts to put LVQ type algorithms on solid mathematical grounds have been made. Remarkably, LVQ-type classifiers fall into the class of *large-margin algorithms* which allow for dimensionality independent generalization bounds, as pointed out first by Crammer et al. (2003). Here, the term margin refers to the so-called hypothesis margin of an LVQ classifier: the distance the classification boundary, that is, the respective prototypes, can be moved without changing the classification of the given data points. Similar bounds can also be derived for recent variants of LVQ which substitute the standard Euclidean metric by an alternative adaptive choice involving

relevance factors (Hammer et al., 2005a; Hammer and Villmann, 2002). Thus, LVQ type networks seem promising candidates for the classification of high-dimensional data sets.

However, standard LVQ training does not directly aim at an optimization of the margin or even an optimization of the classification error, usually. Rather, the learning algorithm is often purely heuristically motivated and does not directly follow any learning objective or explicit cost function. Apart from the fact that the quality of generalization in these methods is therefore not clear, often dynamical problems such as divergence of the algorithms can be observed. An important example is LVQ2.1 and similar strategies which can display divergent behavior unless a proper (also heuristic) window rule is included in the update scheme. Few approaches try to cope with this problem by explicitly grounding the update scheme on a cost function (see, for example, Bottou, 1991). A prominent example has been proposed by Sato and Yamada (1995) together with a discussion of the stability (Sato and Yamada, 1998), however, without considering the borders of receptive fields. A generalization which allows the incorporation of more general metrics, accompanied by an investigation of the borders, has been presented by Hammer and Villmann (2002) and Hammer et al. (2005c). Recently, two models which provide a cost function by means of soft assignments have been proposed by Seo et al. (2003) and Seo and Obermayer (2003). While the first method does not possess a crisp limit, the second one does. However, only the cost function discussed by Sato and Yamada (1995) and Hammer and Villmann (2002) is directly connected to the hypothesis margin, as pointed out by Hammer et al. (2005a). For all these cost functions, the connection to the classification error is not obvious. Thus, there is a clear need for a rigorous analysis of LVQ type learning schemes which allows for a systematic judgement of their dynamical properties and generalization ability.

In this work we discuss and employ a theoretical framework which makes possible the systematic investigation and comparison of LVQ training procedures. The analysis is performed for model situations in which training uses a sequence of uncorrelated, randomized example data. In the theory of such on-line learning processes, concepts known from statistical physics have been applied with great success in the context of supervised and unsupervised learning. This includes perceptron training, gradient based training of layered feedforward neural networks, and unsupervised clustering or principal component analysis (see Watkin et al., 1993; Biehl and Caticha, 2003; Saad, 1999; Engel and van den Broeck, 2001, for additional references). Here, we transfer these methods to LVQ type learning, and we are capable of presenting quite unexpected insights into the dynamical behavior of several important learning schemes.

The essential ingredients of the approach are (1) the consideration of very large systems in the so-called thermodynamic limit and (2) the performing of averages over the randomness or *disorder* contained in the data. A small number of characteristic quantities is sufficient to describe the essential properties of very large systems. Under simplifying assumptions, the evolution of these macroscopic *order parameters* is given by a system of deterministic coupled ordinary differential equations (ODE) which describe the dynamics of on-line learning exactly in the thermodynamic limit.

The formalism enables us to calculate, for instance, typical learning curves exactly within the framework of model situations. We compare the generalization behavior of several LVQ algorithms. Their asymptotic behavior in the limit of infinitely many examples is of particular relevance in this context. We evaluate the achievable generalization error as a function of the prior frequencies of classes and other model parameters. Thus, for the considered model situation, we can rigorously compare the performance of LVQ-type learning schemes and identify the best training method.

The paper is organized as follows: In the next section we introduce the model data and the structure of the LVQ classifier. We also describe the considered training algorithms which include Kohonen's LVQ1 and a limit version of LVQ2.1 or LVQ-LR. We will use the term LVQ+/- for this prescription and we consider it with and without an idealized *early stopping* procedure. Furthermore we study a *crisp* version of RSLVQ (Seo and Obermayer, 2003), which will be referred to as learning from mistakes (LFM). Finally, we consider LFM-W, a variant of the latter which selects data from a *window* close to the decision boundary and relates to the cost function based LVQ2 algorithm (Kohonen et al., 1988; Bottou, 1991). In Section 3 we outline the mathematical treatment. The formalism and its application to the above mentioned algorithms are detailed in Appendix A. Relevant features of the learning curves, including the achievable generalization ability are discussed and compared for the different training prescriptions in Section 4. Finally, we summarize our findings in Section 5 and conclude with a discussion of prospective projects and extensions of our studies.

Some aspects of this work have been communicated previously, in much lesser detail, at the *Workshop on the Self-Organizing Map* in Paris, 2005 (Ghosh et al., 2005).

2. The Model

In this paper we investigate the dynamics of different LVQ algorithms in the framework of a simplifying model situation: High-dimensional data from a mixture of two overlapping Gaussian clusters are presented to a system of two prototype vectors each of which represents one of the classes. Whereas the model appears very much simplified in comparison with real world multi-class, multi-prototype problems, it provides a setting in which it is well possible to study unexpected, essential, and non-trivial features of the learning scenarios.

2.1 Prototype Vectors

We will investigate situations with only two prototype or codebook vectors $w_S \in \mathbb{R}^N$. Here, the subscript $S = \pm 1$ indicates which class of data the vector is supposed to represent. For the sake of brevity we will frequently use the shorthand subscripts $+$ or $-$ for $+1$ and -1, respectively. The classification as parameterized by the prototypes is based on Euclidean distances: Any given input $\xi \in \mathbb{R}^N$ will be assigned to the class of the closest prototype. In our model situation, the classification result is S whenever $| w_{+S} - \xi | < | w_{-S} - \xi |$.

Note that, in principle, the simple LVQ classifier with only two prototypes could be replaced by a linearly separable classification $S = \mathrm{sign}[w_{perc} \cdot \xi - \theta_{perc}]$ with *perceptron* weight vector $w_{perc} = (w_+ - w_-)$ and threshold $\theta_{perc} = \left(w_+^2 - w_-^2\right)/2$. Here, however, we are interested in the more complex learning dynamics of independent codebook vectors as provided by LVQ algorithms. We will demonstrate that our simple example scenario already displays highly non-trivial features and provides valuable insights in the more general training of several prototypes.

2.2 The Data

Throughout the following we consider random inputs which are generated according to a bimodal distribution. We assume that vectors $\xi \in \mathbb{R}^N$ are drawn independently according to the density

$$P(\xi) = \sum_{\sigma=\pm 1} p_\sigma P(\xi \,|\, \sigma) \;\text{ with }\; P(\xi \,|\, \sigma) = \frac{1}{(2\pi v_\sigma)^{N/2}} \exp\left[-\frac{1}{2v_\sigma}\left(\xi - \lambda B_\sigma\right)^2\right]. \qquad (1)$$

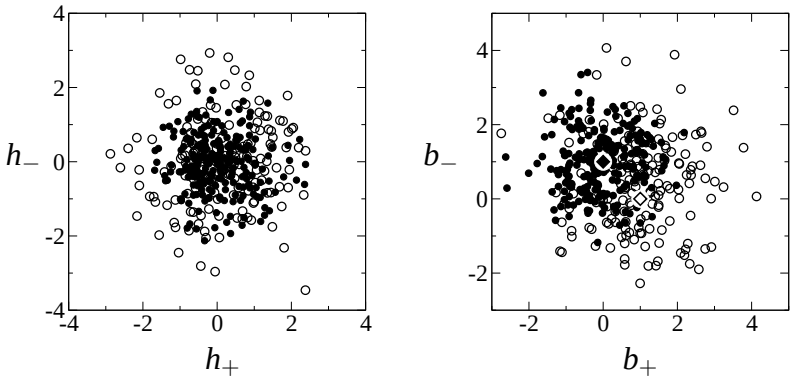

Figure 1: Data as generated according to the density (1) in $N = 200$ dimensions with prior weights $p_- = 0.6, p_+ = 0.4$ and variances $v_- = 0.64, v_+ = 1.44$. Open (filled) circles represent $160\,(240)$ vectors ξ from clusters centered about orthonormal vectors λB_+ (λB_-) with $\lambda = 1$, respectively. The left panel shows the projections $h_\pm = w_\pm \cdot \xi$ of the data on a randomly chosen pair of orthogonal unit vectors $w_\pm$. The right panel corresponds to the projections $b_\pm = B_\pm \cdot \xi$, diamonds mark the position of the cluster centers.

The conditional densities $P(\xi \mid \sigma = \pm 1)$ correspond to isotropic Gaussian clusters with variances v_σ centered at λB_σ. The cluster weights or prior probabilities p_σ satisfy the condition $p_+ + p_- = 1$. We assume the center vectors B_σ to be orthonormal, that is, $B_+^2 = B_-^2 = 1$ and $B_+ \cdot B_- = 0$ in the following. While the distance of the cluster centers is controlled by the model parameter $\lambda = O(1)$, the orthogonality condition fixes their position with respect to the origin. The target classification is taken to coincide with the cluster label $\sigma = \pm 1$. Note that, due to the overlap of the clusters, the task is obviously not linearly separable.

Statistical physics has been used earlier to analyse the generalization behavior of perceptron type learning prescriptions in similar models. We refer to the work of Barkai et al. (1993) and Meir (1995) for examples. The former considers learning from data in the limit of large separations $\lambda \gg v_+, v_-$. The latter compares the off-line minimization of the training error with a maximum-likelihood estimation of the cluster centers. The learning of a linearly separable rule which is defined for clustered input data has been studied in the work of Marangi et al. (1995) and Riegler et al. (1996). Here we are interested in the dynamics and generalization ability of LVQ-type learning rules for non-separable data.

In the following we denote conditional averages over $P(\xi \mid \sigma)$ by $\langle \cdots \rangle_\sigma$ whereas $\langle \cdots \rangle = \sum_{\sigma=\pm 1} p_\sigma \langle \cdots \rangle_\sigma$ represents mean values with respect to the full density (1). According to Eq. (1), the components ξ_j are statistically independent quantities with variance v_σ. For an input from cluster σ we have, for instance, $\langle \xi_j \rangle_\sigma = \lambda (B_\sigma)_j$ and hence

$$\begin{aligned} \langle \xi^2 \rangle_\sigma &= \sum_{j=1}^{N} \langle \xi_j^2 \rangle_\sigma = \sum_{j=1}^{N} \left(v_\sigma + \langle \xi_j \rangle_\sigma^2 \right) = v_\sigma N + \lambda^2 \\ \Rightarrow \langle \xi^2 \rangle &= (p_+ v_+ + p_- v_-)\, N + \lambda^2 . \end{aligned} \tag{2}$$

In the mathematical treatment we will exploit formally the thermodynamic limit $N \to \infty$, corresponding to very high-dimensional data and prototypes. Among other simplifying consequences this allows, for instance, to neglect the term λ^2 on the right hand side of Eq. (2).

The clusters overlap significantly and, moreover, their separation is along a single direction $(B_+ - B_-)$ in the N-dimensional input space. Figure 1 displays sample data in $N = 200$ dimensions, generated according to a density of the form (1). While the clusters are clearly visible in the plane spanned by B_+ and B_-, projections into a randomly chosen two-dimensional subspace do not display the separation at all.

2.3 On-line Algorithms

We consider so-called on-line learning schemes in which a sequence of single example data $\{\xi^\mu, \sigma^\mu\}$ is presented to the system. At each learning step, the update of prototype vectors is based only on the current example. Consequently, an explicit storage of data in the learning system is avoided and the computational effort per step is low compared to costly off-line or batch prescriptions.

In many practical situations examples are drawn from a given training set with possible multiple presentations of the same data. Here, however, we will assume that at each time step $\mu = 1, 2, \ldots$ a new, uncorrelated vector ξ^μ is generated independently according to the density (1). We will treat incremental updates of the generic form

$$w_S^\mu = w_S^{\mu-1} + \Delta w_S^\mu \;\text{ with }\; \Delta w_S^\mu = \frac{\eta}{N} f_S\left[d_+^\mu, d_-^\mu, \sigma^\mu, \ldots\right]\left(\xi^\mu - w_S^{\mu-1}\right) \tag{3}$$

where the vector w_S^μ denotes the prototype after presentation of μ examples and the constant learning rate η is rescaled with the input dimension N.

Prototypes are always moved towards or away from the current input ξ^μ. The modulation function $f_S[\ldots]$ in Eq. (3) defines the actual training algorithm and controls the sign and magnitude of the update of vectors w_S. In general, f_S will depend on the prototype label S and the class label σ^μ of the example; for convenience we denote the dependence on S as a subscript rather than as an explicit argument. The modulation function can further depend on the squared Euclidean distances of ξ^μ from the current positions of the prototypes:

$$d_S^\mu = \left(\xi^\mu - w_S^{\mu-1}\right)^2.$$

Additional arguments of f_S, for instance the lengths or relative positions of the vectors $w_S^{\mu-1}$, could be introduced easily.

The use of a different learning rate per class or even per prototype can be advantageous in practice. This is expected, in particular, if the training data is highly unbalanced with respect to the class membership of examples. Here, we restrict the analysis to using a unique value of η, for simplicity. As we will see, the behavior of algorithms is qualitatively the same in large ranges of prior weights $p_\pm$.

In this work we study the performance of several on-line training prescriptions which can be written in the form of Eq. (3):

a) LVQ1

Kohonen's original formulation of learning vector quantization (Kohonen, 1995, 1997) extends the concept of unsupervised competitive learning in an intuitive way to a classification

task. At each learning step, the prototype with minimal distance from the current example is determined. Only this so-called *winner* is updated, hence the term *winner-takes-all* (WTA) algorithms has been coined for this type of prescription. The winner is moved towards (away from) the example input ξ^μ if prototype label and class membership of the data agree (differ), respectively.
This plausible strategy realizes a compromise between (I) the representation of data by prototypes with the same class label and (II) the identification of decision boundaries which clearly separate the different classes.
LVQ1 for two prototypes is defined by Eq. (3) with the modulation function

$$f_S[d_+^\mu, d_-^\mu, \sigma^\mu] = \Theta\left(d_{-S}^\mu - d_{+S}^\mu\right) S\sigma^\mu \quad \text{with} \quad \Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else.} \end{cases} \tag{4}$$

Here, the Heaviside function singles out the winning prototype and the factor $S\sigma^\mu = \pm 1$ determines the sign of the update. Eq. (4) expands to $f_+ = \Theta(d_-^\mu - d_+^\mu)\sigma^\mu$ for the prototype representing class $S = +1$ (or '+', for short). Correspondingly, it reads $f_- = -\Theta(d_+^\mu - d_-^\mu)\sigma^\mu$ for prototype label $S = -1$.

Note that LVQ1 cannot be interpreted as a stochastic gradient descent procedure in a straightforward way. This is in contrast to unsupervised vector quantization (VQ), which disregards labels completely and would reduce to the choice $f_S[\ldots] = \Theta\left(d_{-S}^\mu - d_{+S}^\mu\right)$ in our model situation (Biehl et al., 1997, 2005; Ghosh et al., 2005).

b) LVQ+/-

A popular modification of basic LVQ was also suggested by Kohonen (1997, 1990) and was termed LVQ2.1. Originally it is based on heuristics and aims at a more efficient separation of prototypes which represent different classes. Given a single example, the update concerns two prototypes: (I) the closest among all codebook vectors which belong to the same class as the data, and (II) the closest vector among the prototypes that represent a different class. The so-called *correct winner* (I) is moved towards the data whereas the *wrong winner* (II) is moved even farther away.

An important ingredient of LVQ2.1 is that the update is only performed if the example ξ^μ falls into the vicinity of the current decision boundary. Variants have been suggested, such as LVQ-LR (Seo and Obermayer, 2003), which differ in the way this selective *window* is defined.

In order to obtain a first insight into this type of algorithm, we consider the unrestricted version or, in other words, the limit of infinite window size. We will refer to this basic prescription as LVQ+/-, in the following.

In our model with two prototypes only, the identification of the respective winners is redundant and the prescription reduces to updating both vectors from each example. The modulation function in Eq. (3) merely determines the sign of Δw_S according to the class label:

$$f_S[\sigma^\mu] = S\sigma^\mu = \pm 1. \tag{5}$$

Algorithms like LVQ2.1 or LVQ-LR were suggested in order to improve the achieved generalization ability in comparison with basic LVQ1. It is well known, however, that a suitable

window size has to be chosen in order to avoid divergence and stability problems. We will demonstrate for unrestricted LVQ+/- that these difficulties occur already in our simplifying model scenario. Various strategies have been suggested in the literature to cope with this problem while keeping the basic form of the algorithm. Here, we will focus on a so-called *early stopping* strategy, which aims at ending the learning process as soon as the generalization ability begins to deteriorate due to the instability.

c) LFM

Recently, Seo and Obermayer suggested a cost function for LVQ which is based on *likelihood ratios* (Seo and Obermayer, 2003). They derive the corresponding *robust soft learning vector quantization* (RSLVQ) scheme which is supposed to overcome the above mentioned stability problems of LVQ+/-. It has been demonstrated that RSLVQ has the potential to yield very good generalization behavior in practical situations (Seo and Obermayer, 2003).

The *crisp* version of RSLVQ is particularly transparent: In the limiting case of deterministic assignments an update analogous to LVQ+/- is performed only if the current configuration of prototypes misclassifies the new example. If, on the contrary, the correct winner is indeed the closest of all prototypes, the configuration is left unchanged. In analogy to perceptron training (e.g., Watkin et al., 1993; Engel and van den Broeck, 2001), we use the term *learning from mistakes* (LFM) for this prescription.

In our model scenario LFM is implemented by the choice

$$f_S[d_+^\mu, d_-^\mu, \sigma^\mu] = \Theta\left(d_{\sigma^\mu}^\mu - d_{-\sigma^\mu}^\mu\right) S\sigma^\mu. \tag{6}$$

Here, the Heaviside function restricts the update to examples which are misclassified upon presentation.

d) LFM-W

The LFM scheme (6) can also be obtained as a limiting case of Kohonen's LVQ2 algorithm, which includes a window rule for the selection of data (Kohonen et al., 1988). As shown by Bottou (1991), this prescription can be interpreted as a stochastic gradient descent with respect to a proper cost function.

As in LFM, an update of the *correct winner* w_{σ^μ} and the *wrong winner* $w_{-\sigma^\mu}$ is only performed if the current classification is wrong, that is, $d_{-\sigma^\mu}^\mu < d_{\sigma^\mu}^\mu$ for an example from class σ^μ. Here, however, the additional constraint

$$\left(\xi^\mu - w_{\sigma^\mu}\right)^2 - \left(\xi^\mu - w_{-\sigma^\mu}\right)^2 \quad < \quad c\left(\xi^\mu - w_{-\sigma^\mu}\right)^2 \tag{7}$$

has to be satisfied, where c is a small positive number (Kohonen et al., 1988; Bottou, 1991). Hence, the vector ξ^μ is only accepted for update if it is not too far away from the correct winner. In other words, ξ^μ has to be misclassified and fall into a window in the vicinity of the current decision boundary.

For the type of high-dimensional data we consider here, the window size c has to satisfy a particular scaling to be meaningful. While the term $(\xi^\mu)^2 = O(N)$ cancels on the left hand side (l.h.s.) of Eq. (7), it dominates the r.h.s. in the limit $N \to \infty$ and the condition is non-trivial only if $c = O(1/N)$. Hence, we introduce a rescaled parameter $\delta = c\,(\xi^\mu)^2 = O(1)$ and obtain $0 < (d_{\sigma^\mu}^\mu - d_{-\sigma^\mu}^\mu) < \delta$ as the conditions for non-zero updates. The first one corresponds to the misclassification of ξ^μ, the second implements the window selection in the limit $N \to \infty$.

We refer to this algorithm as LFM with window (LFM-W) in the following. It is represented by the modulation function

$$f_S[d_+^\mu, d_-^\mu, \sigma^\mu] = \left[\Theta\left(d_{\sigma^\mu}^\mu - d_{-\sigma^\mu}^\mu\right) - \Theta\left(d_{\sigma^\mu}^\mu - d_{-\sigma^\mu}^\mu - \delta\right)\right] S\sigma^\mu. \tag{8}$$

Unrestricted LFM (6) is, of course, recovered in the limit of infinite window size $\delta \to \infty$.

2.4 Relevance of the Model

Before we proceed with the mathematical analysis, we would like to put our model into perspective and discuss its relevance.

Obviously, the learning scenario under consideration is very much simplifying and clear cut. It represents perhaps the simplest non-trivial situation in which LVQ like learning should be applicable. The presence of only one, spherically symmetric Gaussian cluster of data per class is certainly far away from practical situations encountered in typical applications. The use of only two prototype vectors is perfectly appropriate in such a scenario, and the problem of model selection is completely avoided.

Nevertheless, our model represents an important aspect of more realistic multi-class multi-prototype situations: the competition of the two prototypes which currently define the decision boundary or a portion thereof. Note that many practical algorithms reduce to the modification of only one or two prototype vectors in every single update step. Note furthermore that the model also allows to investigate unsupervised vector quantization (VQ), which is equivalent to the competition of two prototypes within the same class in our framework (see, for example, Biehl et al., 1997 and Ghosh et al., 2004).

As we will demonstrate in the following, highly non-trivial properties of the considered training algorithms will become apparent in the course of our analysis. While additional phenomena and complications must be expected in more complex training scenarios, the effects observed here will certainly persist and play an important role under more general circumstances.

Some of the training algorithms we consider are simplifying limit cases of prescriptions suggested in the literature. For example, the above defined LFM algorithm can be interpreted as the crisp limit of RSLVQ (Seo and Obermayer, 2003). Obviously, results for the former will not apply immediately to the latter. Nevertheless, we aim at a principled understanding of how certain ingredients of an algorithm influence its learning dynamics and generalization behavior. The ultimate goal is of course to exploit this understanding in the optimization of existing schemes and the development of novel, efficient ones.

3. Mathematical Analysis

We apply the successful theory of on-line learning (see, for example, Biehl and Caticha, 2003; Engel and van den Broeck, 2001; Saad, 1999, for reviews) to describe the dynamics of LVQ algorithms. The mathematical treatment of our model is based on the assumption of high-dimensional data and prototypes $\xi, w_\pm \in \mathbb{R}^N$. The thermodynamic limit $N \to \infty$ allows to apply concepts from statistical physics in the following key steps which will be detailed below:

1. The original system including many degrees of freedom is characterized in terms of only a few quantities, so-called *macroscopic order parameters.* For these, recursion relations can be derived from the learning algorithm.

2. Applying the central limit theorem enables us to perform an average over the random sequence of example data by means of Gaussian integrations.

3. Self-averaging properties of the order parameters allow to restrict the description to their mean values. Fluctuations of the stochastic dynamics can be neglected in the limit $N \to \infty$.

4. A *continuous time limit* leads to the description of the dynamics in terms of coupled, deterministic ordinary differential equations (ODE) for the above mentioned order parameters.

5. The (numerical) integration of the ODE for given modulation function and initial conditions yields the evolution of order parameters in the course of learning. From the latter one can directly obtain the learning curve, that is, the generalization ability of the LVQ classifier as a function of the number of example data.

3.1 Characteristic Quantities and Recursions

The selection of meaningful macroscopic quantities reflects, of course, the particular structure of the model. After presentation of μ examples, the system is characterized by the high-dimensional vectors w_+^μ, w_-^μ and their positions relative to the center vectors B_+, B_-. As we will demonstrate in the following, a suitable set of order parameters is

$$R_{S\sigma}^\mu = w_S^\mu \cdot B_\sigma \quad \text{and} \quad Q_{ST}^\mu = w_S^\mu \cdot w_T^\mu \quad \text{with} \quad \sigma, S, T \in \{-1, +1\}.$$

The self-overlaps Q_{++}^μ, Q_{--}^μ and the symmetric cross-overlap $Q_{+-}^\mu = Q_{-+}^\mu$ relate to the lengths and relative angle between the prototype vectors, whereas the four quantities $R_{S\sigma}^\mu$ specify their projections into the subspace spanned by $\{B_+, B_-\}$.

From the generic algorithm (3) we can directly derive recursion relations which specify the change of order parameters upon presentation of example ξ^μ from cluster σ^μ:

$$\begin{aligned}
\frac{R_{S\sigma}^\mu - R_{S\sigma}^{\mu-1}}{1/N} &= \eta\, f_S \left(B_\sigma \cdot \xi^\mu - R_{S\sigma}^{\mu-1} \right), \\
\frac{Q_{ST}^\mu - Q_{ST}^{\mu-1}}{1/N} &= \eta\, f_S \left(w_T^{\mu-1} \cdot \xi^\mu - Q_{ST}^{\mu-1} \right) + \eta\, f_T \left(w_S^{\mu-1} \cdot \xi^\mu - Q_{ST}^{\mu-1} \right) \\
&\quad + \eta^2\, f_S\, f_T\, (\xi^\mu)^2 / N. \qquad (9)
\end{aligned}$$

Here we use the shorthand f_S for the modulation function of prototype S, omitting the arguments.

3.2 Average over Random Examples

For a large class of modulation functions, including the ones considered here, the current input ξ^μ appears on the right hand side of Eq. (9) only through its length and the projections

$$h_S^\mu = w_S^{\mu-1} \cdot \xi^\mu \ \text{ and } \ b_\sigma^\mu = B_\sigma \cdot \xi^\mu. \qquad (10)$$

Note that also the Heaviside terms as they appear in the modulation functions, Eqs. (4,6,8), do not depend on ξ^μ explicitly, for example:

$$\Theta\left(d_-^\mu - d_+^\mu\right) = \Theta\left(+2(h_+^\mu - h_-^\mu) - Q_{++}^{\mu-1} + Q_{--}^{\mu-1}\right).$$

When performing the average over the current example ξ^μ we first exploit Eq. (2) which yields

$$\lim_{N\to\infty} \frac{1}{N} \left\langle (\xi^\mu)^2 \right\rangle = (v_+ p_+ + v_- p_-)$$

for all input vectors in the thermodynamic limit.

We assume that the new example input ξ^μ is uncorrelated with all previous data and, thus, also with the current vectors $w_S^{\mu-1}$. Therefore, the central limit theorem implies that the projections, Eq. (10), become correlated Gaussian quantities for large N. Note that this observation does not rely on the fact that the specific density (1) is a mixture of Gaussians itself. It holds whenever the components of ξ^μ are statistically independent and have the same class conditional first and second moments as in (1).

The joint Gaussian density $P(h_+^\mu, h_-^\mu, b_+^\mu, b_-^\mu)$ is fully specified by first and second moments. As shown in the appendix, for an input from cluster σ these read

$$\left\langle h_S^\mu \right\rangle_\sigma = \lambda R_{S\sigma}^{\mu-1}, \quad \left\langle b_\tau^\mu \right\rangle_\sigma = \lambda \delta_{S\tau}, \quad \left\langle h_S^\mu h_T^\mu \right\rangle_\sigma - \left\langle h_S^\mu \right\rangle_\sigma \left\langle h_T^\mu \right\rangle_\sigma = v_\sigma Q_{ST}^{\mu-1},$$

$$\left\langle h_S^\mu b_\tau^\mu \right\rangle_\sigma - \left\langle h_S^\mu \right\rangle_\sigma \left\langle b_\tau^\mu \right\rangle_\sigma = v_\sigma R_{S\tau}^{\mu-1}, \quad \left\langle b_\rho^\mu b_\tau^\mu \right\rangle_\sigma - \left\langle b_\rho^\mu \right\rangle_\sigma \left\langle b_\tau^\mu \right\rangle_\sigma = v_\sigma \delta_{\rho\tau} \tag{11}$$

where $S, T, \sigma, \rho, \tau \in \{+1, -1\}$ and $\delta_{\ldots}$ is the Kronecker-Delta. Hence, the density of $h_\pm^\mu$ and $b_\pm^\mu$ is given in terms of the model parameters $\lambda, p_\pm, v_\pm$ and the above defined set of order parameters in the previous time step.

This important result enables us to perform an average of the recursion relations, Eq. (9), over the latest example data in terms of Gaussian integrations. Moreover, the result can be expressed in closed form in $\{R_{S\sigma}^{\mu-1}, Q_{ST}^{\mu-1}\}$. In the appendix we detail the calculation and give specific results for the algorithms considered here.

3.3 Self-averaging Property

The thermodynamic limit has an additional simplifying consequence: Fluctuations of the quantities $\{R_{S\sigma}^\mu, Q_{ST}^\mu\}$ decrease with increasing N and in fact vanish for $N \to \infty$. Hence, a description in terms of their mean values is sufficient. In the statistical physics of disordered systems the term *self-averaging* is used for this property.

For a detailed mathematical discussion of self-averaging in on-line learning we refer to the work of Reents and Urbanczik (1998). Here, we have confirmed this property empirically in Monte Carlo simulations of the learning process by comparing results for various values of N, see Figure 2 (right panel) and the discussion in Section 3.6.

Neglecting fluctuations allows us to interpret the averaged recursion relations directly as recursions for the means of $\{R_{S\sigma}^\mu, Q_{ST}^\mu\}$ which coincide with their actual values in very large systems.

3.4 Continuous Learning Time

An essentially technical simplification is due to the fact that, for $N \to \infty$, we can interpret the differences on the left hand sides of Eq. (9) as derivatives with respect to the continuous learning time

$$\alpha = \mu/N.$$

Clearly, one would expect that the number of examples required for successful training should grow linearly with the number of adjustable parameters in the system. Hence, the rescaling of μ with the dimension of the feature and prototype space appears natural.

The resulting set of coupled ODE obtained from Eq. (9) is of the form

$$\begin{aligned}\frac{dR_{S\tau}}{d\alpha} &= \eta\left(\langle b_\tau f_S\rangle - R_{S\tau}\langle f_S\rangle\right),\\ \frac{dQ_{ST}}{d\alpha} &= \eta\left(\langle h_S f_T + h_T f_S\rangle - Q_{ST}\langle f_S + f_T\rangle\right)\\ &\quad +\eta^2\sum_{\sigma=\pm1} v_\sigma p_\sigma \langle f_S f_T\rangle_\sigma .\end{aligned} \tag{12}$$

The required averages and specific equations for the particular modulation functions considered here are given in the Appendix, Sections A.3 and A.4. Note that the ODE for Q_{ST} contain terms proportional to η and η^2 while $dR_{S\tau}/d\alpha$ is linear in the learning rate.

3.5 The Learning Curve

After working out the system of ODE for a specific modulation function, it can be integrated, at least numerically. For given initial conditions $\{R_{S\sigma}(0), Q_{ST}(0)\}$, learning rate η, and model parameters $\{p_\pm, v_\pm, \lambda\}$ one obtains the typical evolution $\{R_{S\sigma}(\alpha), Q_{ST}(\alpha)\}$. This is the basis for analysing the performance and convergence properties of various algorithms in the following.

Most frequently, we will consider prototypes that are initialized as independent random vectors of squared length $\widehat{Q}$ with no prior knowledge about the cluster positions. In terms of order parameters this implies in our model

$$Q_{++}(0) = Q_{--}(0) = \widehat{Q},\;\; Q_{+-}(0) = 0 \;\;\text{ and }\; R_{S\sigma}(0) = 0 \;\;\text{ for all }\; S,\sigma = \pm 1. \tag{13}$$

Obviously, the precise initial positions of prototypes with respect to the cluster geometry can play a crucial role for the learning dynamics. In the next section we will demonstrate that the outcome of, for instance, LFM and LVQ+/- with early stopping can depend strongly on the initial positions of prototypes. On the contrary, asymptotic properties of, for example, LVQ1 training will prove independent of initialization in the limit of infinitely long training sequences within our model situation.

After training, the success of learning can be quantified in terms of the generalization error, that is, the probability for misclassifying novel, random data which did not appear in the training sequence. Here we can consider the two contributions for misclassifying data from cluster $\sigma = 1$ or $\sigma = -1$ separately:

$$\varepsilon = p_+\varepsilon_+ + p_-\varepsilon_- \quad \text{with} \;\; \varepsilon_\sigma = \langle \Theta(d_{+\sigma} - d_{-\sigma})\rangle_\sigma . \tag{14}$$

Exploiting the central limit theorem in the same fashion as above, one can work out the generalization error as an explicit function of the order parameters. As detailed in the appendix, Section A.6, one obtains for the above contributions $\varepsilon_\pm$:

$$\varepsilon_\sigma = \Phi\left(\frac{Q_{\sigma\sigma} - Q_{-\sigma-\sigma} - 2\lambda(R_{\sigma\sigma} - R_{-\sigma\sigma})}{2\sqrt{v_\sigma}\sqrt{Q_{++} - 2Q_{+-} + Q_{--}}}\right) \quad \text{where} \;\; \Phi(z) = \int_{-\infty}^{z} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}}. \tag{15}$$

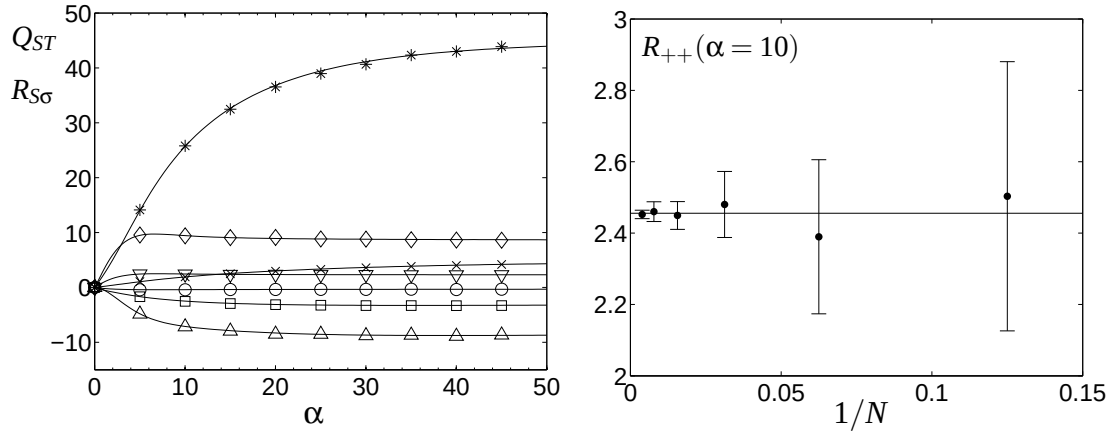

Figure 2: LVQ1 with $\lambda = 2$, $v_+ = 4, v_- = 9$, $p_+ = 0.8$, and learning rate $\eta = 1.0$.
Left panel: Characteristic overlaps vs. $\alpha = \mu/N$. Solid lines display the result of integrating the system of ODE for initial conditions as in Eq. (13) with $\widehat{Q} = 10^{-4}$. Symbols represent Monte Carlo simulations for $N = 100$, on average over 100 independent runs. Standard error bars would be smaller than the symbol size. Curves and symbols correspond to: $Q_{--}\,(*)$, $Q_{++}\,(\Diamond)$, $R_{--}\,(\times)$, $R_{++}\,(\bigtriangledown)$, $R_{+-}\,(\bigcirc)$, $R_{-+}\,(\square)$, and $Q_{+-}\,(\bigtriangleup)$.
Right panel: Example for the self-averaging behavior of order parameters as observed in Monte Carlo simulations for $N = 8, 16, 32, 64, 128,$ and 256. Dots mark the observed average value of $R_{++}(\alpha = 10)$ vs. $1/N$, bars represent the corresponding variance. Here, the latter vanishes approximately like $1/N$, while the mean values display no systematic dependence on N for large enough systems. The horizontal line marks the theoretical prediction for $N \to \infty$.

By inserting $\{R_{S\sigma}(\alpha), Q_{ST}(\alpha)\}$ we obtain the learning curve $\varepsilon_g(\alpha)$, that is, the typical generalization error after on-line training with αN random examples. Here, we once more exploit the fact that the order parameters and, thus, also ε_g are self-averaging, non-fluctuating quantities in the thermodynamic limit $N \to \infty$.

A classification scheme based on two prototypes is restricted to linear decision boundaries. We will therefore compare the performance of LVQ algorithms with the best linear decision (bld) boundary for given parameters $p_\pm, v_\pm$, and λ. For symmetry reasons it is given by a plane orthogonal to $(B_+ - B_-)$. It is straightforward to obtain the corresponding generalization error ε_g^{bld} from Eqs. (14,15) by appropriate parameterization of the plane and minimization of ε_g (Ghosh et al., 2004).

Note that the Bayes optimal classification of data from density (1) is, in general, given by a non-linear decision boundary which contains the vectors ξ with $p_+ P(\xi|+1) = p_- P(\xi|-1)$ (Duda et al., 2000). Only for $v_+ = v_-$ it becomes a plane and LVQ with two prototypes could potentially implement Bayes optimal generalization.

3.6 Restrictions of the Analysis

While we have already discussed the relevance of our simplifying model scenario in Subsection 2.4, we want to summarize here some restrictions of the mathematical analysis.

Perhaps, the consideration of the *thermodynamic limit* $N \to \infty$ of infinite-dimensional feature vectors appears to be the most severe limitation in the mathematical treatment. Together with the assumption of statistically independent features ξ_j^μ, it facilitates the above explained steps of the analysis, the evaluation of averages over random data being the most important one.

We find that our results describe very well the (mean) behavior of systems with a fairly small number of input dimensions, yielding excellent agreement for $N = 100$ already. Fig. 2 (right panel) shows a comparison of different system sizes and illustration of the above mentioned *self-averaging* property. In other learning problems, the behavior of low-dimensional systems may differ significantly from results obtained in the limit $N \to \infty$, as fluctuations which were neglected here become more important. As an example, we mention the symmetry breaking specialization of prototypes in unsupervised VQ (Biehl et al., 1997). Here, however, the self-averaging behavior of order parameters is reflected by the fact that their variances vanish rapidly with N, approximately like $1/N$, see Figure 2 (right panel). At the same time, no systematic dependence of the means on the system size is observed. Hence, our treatment yields an excellent approximation for systems of fairly small dimension N, already: Deviations of observed and predicted values of characteristic overlaps are expected to be on the order $1/\sqrt{N}$. For analytic results concerning such *finite size corrections* in on-line training see, for instance, Saad (1999) and references therein.

Performing averages over the randomness in the data yields *typical properties* of the system in the considered model situations. This method is different in spirit and thus complements other successful approaches in computational learning theory, where one aims at rigorous bounds on the generalization error without making explicit assumptions about the learning scenario (for examples in the context of LVQ, see Crammer et al., 2003; Hammer et al., 2005a). Such bounds are not necessarily *tight* and can be quite far from the actual behavior observed in practical situations. On the contrary, results obtained in the flavor of statistical physics analysis lack the mathematical rigor of strict bounds and may be sensitive to details of the model assumptions, for example, the statistical properties of the data. As an attractive feature, the approach provides information about the system which goes beyond its generalization ability, such as the learning dynamics and the location of prototypes.

For more detailed discussions of strengths and limitations of our approach we refer to the reviews of, for example, Biehl and Caticha (2003), Watkin et al. (1993), and Engel and van den Broeck (2001).

3.7 Relation to Statistical Physics

The relation to statistical physics is not crucial for what follows and may be considered a merely technical point. Nevertheless, for the interested reader, we would like to make a few remarks in this context.

The analogy is more important in the theory of batch or off-line learning. There, the cost function of training is interpreted as an *energy* and the analysis proceeds along the lines of equilibrium statistical mechanics. For an introduction to these concepts we refer to, for example, Biehl and Caticha (2003), Watkin et al. (1993), and Engel and van den Broeck (2001). Several ideas from

this type of analysis do carry over to the investigation of on-line learning which addresses the non-equilibrium dynamics of learning.

In many physical systems it is impossible, and indeed useless, to keep track of all microscopic degrees of freedom. As an example, consider the positions and velocities of, say, N particles in a gas. Similarly, a magnetic material will contain a number N of atoms each of which contributes one elementary magnetic moment. As N gets very large, say, on the order 10^{23} in condensed matter physics, microscopic details become less important and such systems are successfully described by a fairly small number of macroscopic quantities or order parameters. In the above examples, it may be sufficient to know volume and pressure of the gas or the total magnetization emerging from the collaborative behavior of atoms. Mathematically speaking, so-called phase space integrals over all degrees of freedom can be performed by means of a saddle point approximation for $N \to \infty$, reducing the description to a low-dimensional one. Similarly, the high-dimensional phase space trajectories of dynamical systems can be represented by the temporal evolution of a few macroscopic order parameters. The same idea applies here as we describe the learning dynamics of $2N$ prototype components in terms of a few characteristic overlaps and disregard their microscopic details.

An important branch of statistical physics deals with so-called *disordered systems*, where the interacting degrees of freedom are embedded in a random environment. In the above example this could mean that magnetic atoms are randomly placed within an otherwise non-magnetic material, for instance. The correct analytic treatment of such systems requires sophisticated analytical tools such as the famous *replica trick*, see Watkin et al. (1993) and Engel and van den Broeck (2001) and references therein.

These tools have been developed in the context of disordered magnetic materials, indeed, and have been put forward within the statistical physics of learning. In learning theory, the *disorder* emerges from the random generation of training data. In off-line learning, the cost-function or *energy* is defined for one specific set of data only and the generic, typical behavior is determined by performing the *disorder average* over all possible training sets. This requires the above mentioned replica method or approximation techniques which involve subtle mathematical difficulties, see Watkin et al. (1993) and Engel and van den Broeck (2001).

The situation is slightly more favorable in the framework which we resort to here: In on-line learning at each time step of the process, a novel random example is presented. As a consequence the *disorder average* can be performed step by step in the more elementary fashion outlined above (Biehl and Caticha, 2003; Engel and van den Broeck, 2001).

4. Results

In the following we present our results obtained along the lines of the treatment outlined in Sec. 3. We will compare the typical learning curves of LVQ+/-, an idealized early stopping procedure, LFM, and LFM-W with those of LVQ1. The latter, original formulation of basic LVQ serves as a reference each modification has to compete with.

We will put emphasis on the asymptotic behavior in the limit $\alpha \to \infty$, that is, the generalization error achieved from an arbitrarily large number of examples and its dependence on the model parameters. This asymptotic behavior is of particular relevance for comparing different training algorithms. It can be studied by analysing the stable fixed point configuration of the system of ODE. Note that properties thereof will not depend on initialization or the position of cluster cen-

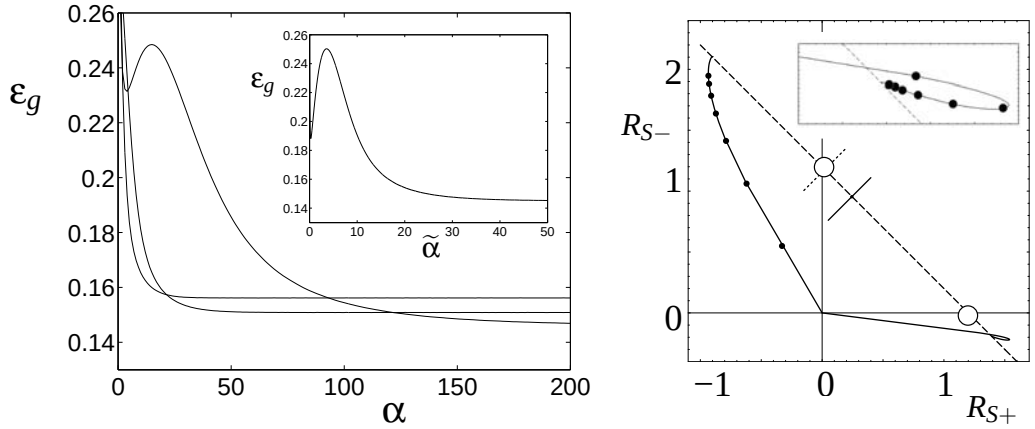

Figure 3: LVQ1 for $\lambda = 1.2, v_+ = v_- = 1$, and $p_+ = 0.8$, initialization as in Figure 2.
Left panel: Learning curves $\varepsilon_g(\alpha)$ for three different learning rates $\eta = 0.2, 1.0, 2.0$ (from bottom to top at $\alpha = 200$). Large η are favorable initially, whereas smaller η yield better asymptotic generalization for $\alpha \to \infty$. The inset shows the limiting behavior for $\eta \to 0$ and $\alpha \to \infty$, that is, ε_g as a function of $\widetilde{\alpha} = \eta\alpha$.
Right panel: Trajectories of prototypes in the limit $\eta \to 0, \alpha \to \infty$. Solid lines correspond to the projections of prototypes into the plane spanned by λB_+ and λB_- (marked by open circles). The dots correspond to the pairs of values $\{R_{S+}, R_{S-}\}$ observed at $\widetilde{\alpha} = \eta\alpha = 2, 4, 6, 8, 10, 12, 14$ in Monte Carlo simulations with $\eta = 0.01$ and $N = 200$, averaged over 100 runs. Note that, because $p_+ > p_-$, w_+ approaches its final position much faster and in fact *overshoots*. The inset displays a close-up of the region around its stationary location. The short solid line marks the asymptotic decision boundary as parameterized by the prototypes, the short dashed line marks the best linear decision boundary. The latter is very close to λB_- for the pronounced dominance of the $\sigma = +1$ cluster with $p_+ = 0.8$.

ters relative to the origin. Asymptotic properties are controlled only by the distance of the centers $|\lambda B_+ - \lambda B_-| = \sqrt{2}\lambda$, the model parameters $v_\pm, p_\pm$ and the learning rate η.

4.1 LVQ1

Figure 2 (left panel) displays the evolution of order parameters in the course of learning for an example choice of the model parameters. Monte Carlo simulations for $N = 100$ already agree very well with the $(N \to \infty)$ theoretical prediction based on integrating the corresponding ODE, Eq. (34) . We consider initialization of prototypes close to the origin, that is, relatively far from the region of high density in our data model. Note that in a WTA algorithm the initial prototypes must not coincide exactly, hence we choose random $w_\pm(0)$ with, for example, squared length $Q_{++} = Q_{--} = \widehat{Q} = 10^{-4}$ in Eq. (13).

The self-averaging property, see Section 3.3, is illustrated in Fig. 2 (right panel). In Monte Carlo simulations one observes that averages of the order parameters over independent runs approach the

theoretical prediction as $N \to \infty$. At the same time, the corresponding variances vanish like $1/N$ with increasing dimension.

The typical learning curve $\varepsilon_g(\alpha)$ is in the center of our interest. Figure 3 (left panel) displays the behavior of LVQ1 for different learning rates in an example situation. The pronounced non-monotonic behavior for smaller η clearly indicates that the prescription is suboptimal: For a range of α, additional information leads to a loss of generalization ability. This effect is particularly pronounced for highly unbalanced data with, say, $p_+ \gg p_-$. The results suggest a schedule which employs large values of η initially and then decreases the learning rate with increasing α. This aspect will be investigated in greater detail in a forthcoming project, here we consider only training with constant η.

For LVQ1 we find that the stationary, asymptotic generalization error $\varepsilon_g^{stat} = \varepsilon_g(\alpha \to \infty)$ decreases with $\eta \to 0$ like

$$\varepsilon_g^{stat}(\eta) = \varepsilon_g^{lvq1} + O(\eta) \quad \text{for small } \eta.$$

Here, ε_g^{lvq1} denotes the best value achievable with LVQ1 for a given set of model parameters. This is analogous to the behavior of stochastic gradient descent procedures like VQ, where the associated cost function is minimized in the simultaneous limits of small learning rates $\eta \to 0$ and $\alpha \to \infty$, such that $\widetilde{\alpha} = \eta\alpha \to \infty$. In absence of a cost function we can still consider this limit. Terms proportional to η^2 can be neglected in the ODE, and the evolution in rescaled learning time $\widetilde{\alpha}$ becomes η-independent. The inset of Fig. 3 (left panel) shows the limiting learning curve $\varepsilon_g(\widetilde{\alpha})$. It displays a strong non-monotonic behavior for small $\widetilde{\alpha}$.

The right panel of Fig. 3 displays the trajectories of prototypes projected into the plane spanned by B_+ and B_-. Note that, as could be expected from symmetry arguments, the $(\alpha \to \infty)$-asymptotic projections of prototypes into the $B_\pm$-plane are along the axis connecting the cluster centers. Moreover, in the limit $\eta \to 0$, their stationary position lies precisely in the plane and fluctuations orthogonal to $B_\pm$ vanish. This is signaled by the fact that the order parameters for $\widetilde{\alpha} \to \infty$ satisfy $Q_{SS} = R_{S+}^2 + R_{S-}^2$, and $Q_{+-} = R_{++}R_{-+} + R_{+-}R_{--}$ which implies

$$w_S(\widetilde{\alpha} \to \infty) = R_{S+}B_+ + R_{S-}B_- \quad \text{for } S = \pm 1. \tag{16}$$

Here we can conclude that the actual prototype vectors approach the above unique configuration, asymptotically. Note that, in general, stationarity of the order parameters does not imply necessarily that $w_\pm$ converge to points in N-dimensional space. For LVQ1 with $\eta > 0$ fluctuations in the space orthogonal to $\{B_+, B_-\}$ persist even for constant $\{R_{S\sigma}, Q_{ST}\}$.

Figure 3 (right) reveals further information about the learning process. When learning from unbalanced data, for example, $p_+ > p_-$ as in the example, the prototype representing the stronger cluster will be updated more frequently and in fact *overshoots,* resulting in the non-monotonic behavior of ε_g. The use of a different learning rate per class could correct this overshooting behavior and recover the transient behavior for equal priors, qualitatively.

The asymptotic ε_g^{lvq1} as achieved by LVQ1 is typically quite close to the potential optimum ε_g^{bld}. Figure 4 displays the asymptotic generalization error as a function of the prior p_+ in two different settings of the model. In the left panel $v_+ = v_-$ whereas the right panel shows an example with different cluster variances. In the completely symmetric situation with equal variances and balanced priors, $p_+ = p_-$, the LVQ1 result coincides with the best linear decision boundary which is through $\lambda(B_+ + B_-)/2$ for this setting. Whenever the cluster-variances are different, the symmetry about $p_+ = 1/2$ is lost but $|\varepsilon_g^{lvq1} - \varepsilon_g^{bld}| = 0$ for one particular (v_+, v_-)-dependent value of $p_+ \in]0,1[$.

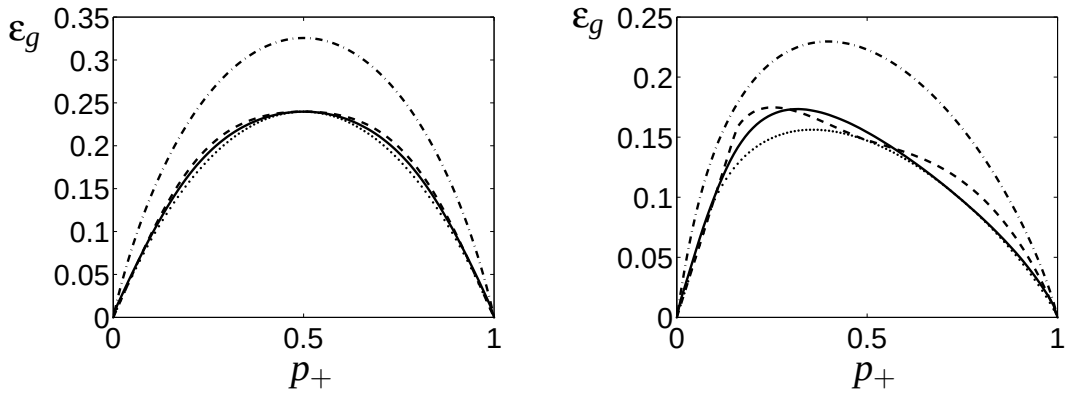

Figure 4: Achievable generalization error in the model with $\lambda = 1$ as a function of p_+. In both panels, the lowest, dotted curve represents ε_g^{bld}, that is, the best linear classifier. Solid lines mark the asymptotic ε_g^{lvq1} of LVQ1, the dashed lines corresponds to ε_g^{stop} as obtained from an idealized early stopping scheme for LVQ+/- with prototypes initialized in the origin. The far from optimal $(\alpha \to \infty)$-asymptotic ε_g^{lfm} result for LFM is marked by chain lines.
Left panel: The case of equal variances, $v_+ = v_- = 1$, ε_g^{lvq1} and ε_g^{stop} coincide for $p_+ = 1/2$ where both are optimal.
Right panel: An example for unequal variances, $v_+ = 0.25, v_- = 0.81$. The result of LVQ+/- with idealized early stopping is still optimal for equal priors, while $\varepsilon_g^{lvq1} = \varepsilon_g^{bld}$ in $p_+ \approx 0.74$ in this setting.

4.2 LVQ+/-

Here we report results concerning the divergent behavior displayed by the LVQ+/- prescription in its basic form. We furthermore show how an appropriate early stopping procedure could overcome this difficulty.

4.2.1 DIVERGENT BEHAVIOR

The structure of the ODE for LVQ+/-, $f_S = S\sigma^\mu$, is particularly simple, see Appendix A.5. Analytic integration yields Eq. (37) for settings with $p_+ \neq p_-$. In this generic case, the evolution of $\{R_{S\sigma}, Q_{ST}\}$ displays a strong divergent behavior: All quantities associated with the prototype representing the weaker cluster display an exponential increase with α.

Figure 5 (left panel) shows an example for $\lambda = 1$, $v_+ = v_- = 1$, learning rate $\eta = 0.5$, and unbalanced priors $p_+ = 0.8, p_- = 0.2$. Here, order parameters which involve w_-, that is, $R_{--}, R_{-+}, Q_{--}, Q_{+-}$, diverge rapidly. On the contrary, quantities which relate only to w_+, that is, R_{++}, R_{+-}, Q_{++}, remain finite and approach well-defined asymptotic values for $\alpha \to \infty$.

This behavior is due to the fact that the *wrong winner* is always moved away from the current data in LVQ+/- training. Clearly, this feature renders the learning dynamics unstable as soon as $p_+ \neq p_-$. Even in our simple model problem, repulsion will always dominate for one of the prototypes and, like w_- in our example, it will be moved arbitrarily far away from the cluster centers as $\alpha \to \infty$.

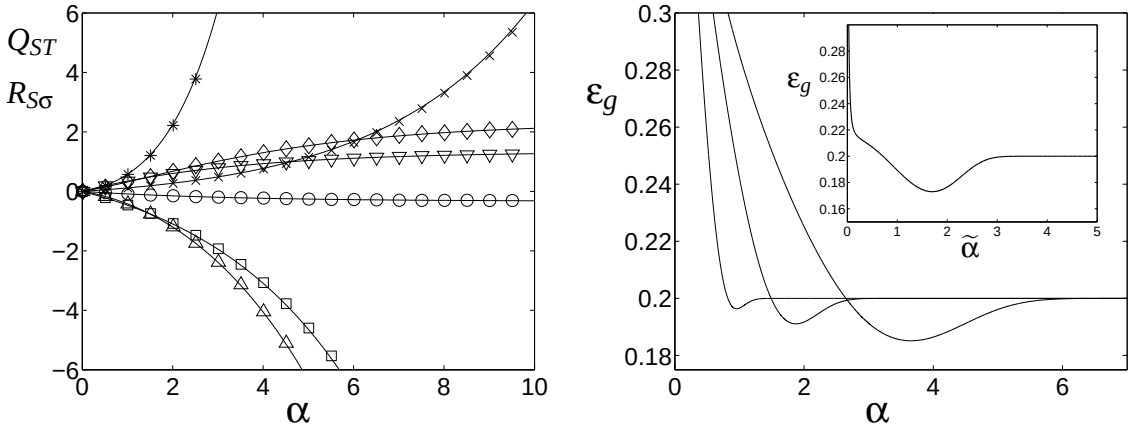

Figure 5: LVQ+/- with $\lambda = 1, v_+ = v_- = 1$, and $p_+ = 0.8$.
Left panel: Characteristic overlaps vs. α for learning rate $\eta = 0.5$. Solid lines correspond to the analytical solution Eq. (37) of the ODE for initialization $w_+(0) = w_-(0) = 0$. Symbols represent Monte Carlo results for $N = 100$ on average over 100 runs. Standard error bars would be smaller than the symbol size. Curves and symbols correspond to Q_{--} ($*$), Q_{++} ($\Diamond$), R_{--} ($\times$), R_{++} ($\bigtriangledown$), R_{+-} ($\bigcirc$), R_{-+} ($\Box$), and Q_{+-} ($\bigtriangleup$).
Right panel: Learning curves $\varepsilon_g(\alpha)$ for three different rates $\eta = 2.0, 1.0$, and 0.5 (from left to right). The generalization error displays a pronounced minimum at intermediate α and approaches the trivial value $\varepsilon_g = \min\{p_+, p_-\}$ asymptotically for $\alpha \to \infty$. The inset shows the asymptotic behavior $\varepsilon_g(\widetilde{\alpha})$ in the simultaneous limit $\eta \to 0$, $\alpha \to \infty$ with rescaled $\widetilde{\alpha} = \eta\alpha$.

The divergent behavior is also apparent in Figure 6, which displays example trajectories of $w_\pm$ in the limit $\eta \to 0$, projected into the $B_\pm$-plane. The projections of prototypes follow straight lines, approximately. Prototype w_+, representing the stronger cluster in this case, approaches a stationary position at the symmetry axis of the density (1). For w_- the less frequent positive updates from the weaker cluster $\sigma = -1$ cannot counterbalance the repulsion.

The resulting classification scheme for $\alpha \to \infty$ is trivial: All data will be assigned to the class of the stronger cluster. Hence, the asymptotic generalization error of unmodified LVQ+/- is given by $\min\{p_+, p_-\}$ in our model, independent of the learning rate η. This can also be seen in the learning curves displayed in Figure 5 (right panel).

Note that the behavior is qualitatively different for the singular case of balanced priors $p_+ = p_- = 1/2$, see Eq. (38) in the appendix for the corresponding analytic solution of the ODE. First, the increase of order parameters with α, which is exponential for $p_+ \neq p_-$, becomes linear ($R_{S\sigma}$) or quadratic (Q_{ST}) in α for equal priors. Second, both prototypes move to infinity, that is, away from the data, as $\alpha \to \infty$. In a visualization of the learning dynamics in the spirit of Figure 6, the trajectories become parallel to the symmetry axis as $p_+ = p_-$. It is interesting to note that, in spite of this divergence, the corresponding decision boundary is optimal in the singular case of equal priors.

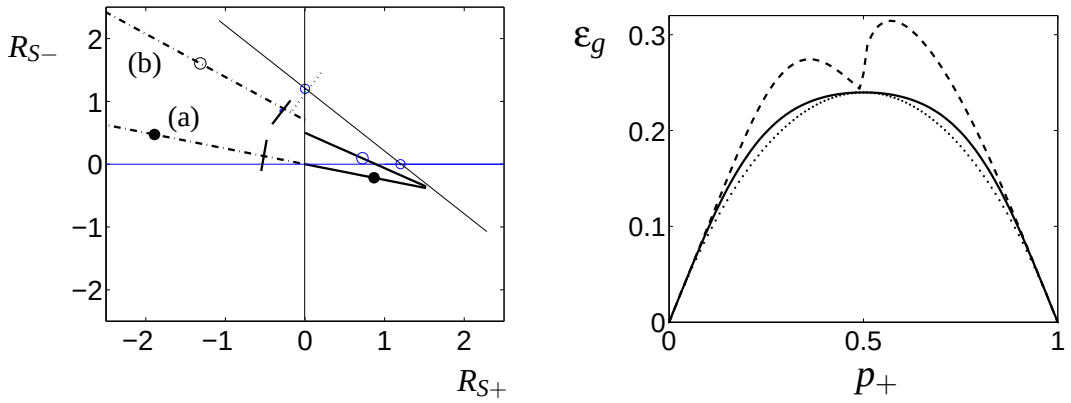

Figure 6: LVQ+/- and early stopping for different initializations in the limit $\eta \to 0, \alpha \to \infty$.
Left panel: Projected trajectories for model parameters $\lambda = 1.2, p_+ = 0.8, v_+ = v_- = 1$. Cluster centers correspond to open circles, the short dotted line marks the best linear decision boundary. Solid (chain) lines show the trace of prototype w_+ (w_-), respectively. Filled symbols display the position after *early stopping* in the minimum of $\varepsilon_g(\widetilde{\alpha})$, short solid lines mark the projection of the respective decision boundaries. Squares correspond to the positions obtained from $w_\pm(0) = 0$ (a), full circles mark the results for initialization (b) with an offset from the origin and the (B_+, B_-)-plane: $R_{++} = R_{-+} = 0, R_{+-} = 0.5, R_{--} = 0.7, Q_{+-} = 0, Q_{++} = 1.8, Q_{--} = 2.9$. In both cases, w_+ approaches the same stationary position on the symmetry axis, while w_- displays divergent behavior as $\widetilde{\alpha} = \eta\alpha$ increases. Note that the decision boundary obtained by early stopping in case (b) appears to be close to optimal in the projection. However, it is tilted strongly in the remaining dimensions as indicated by the violation of condition (16). Consequently, ε_g^{stop} is higher for (b) than it is for initialization (a).
Right panel: Generalization error ε_g^{stop} of idealized early stopping as a function of p_+ for $\lambda = 1$ and equal variances $v_+ = v_- = 1$. The dotted line corresponds to the best linear decision boundary. The solid line marks the outcome of LVQ+/- with early stopping when prototypes are initialized in the origin, case (a) in the left panel, which is also displayed in Fig. 4. The dashed line represents the far from optimal ε_g^{stop} for an example initialization with an offset from the origin, case (b) in the left panel.

4.2.2 EARLY STOPPING

Several heuristic strategies have been suggested in the literature to cure the divergent behavior while keeping the essential ingredients of LVQ+/-:

One possible measure that comes to mind immediately is to let the update of the *wrong winner* depend explicitly on its distance d^μ from the data. The divergence should be much weaker or even seize, provided the magnitude of the update decreases fast enough with d^μ or if it is cut off at a maximum distance d_{max}.

Another popular strategy is to update only from data that falls into a *window* close to the current decision boundary, that is, close to the midplane between the prototypes (Kohonen, 1997).

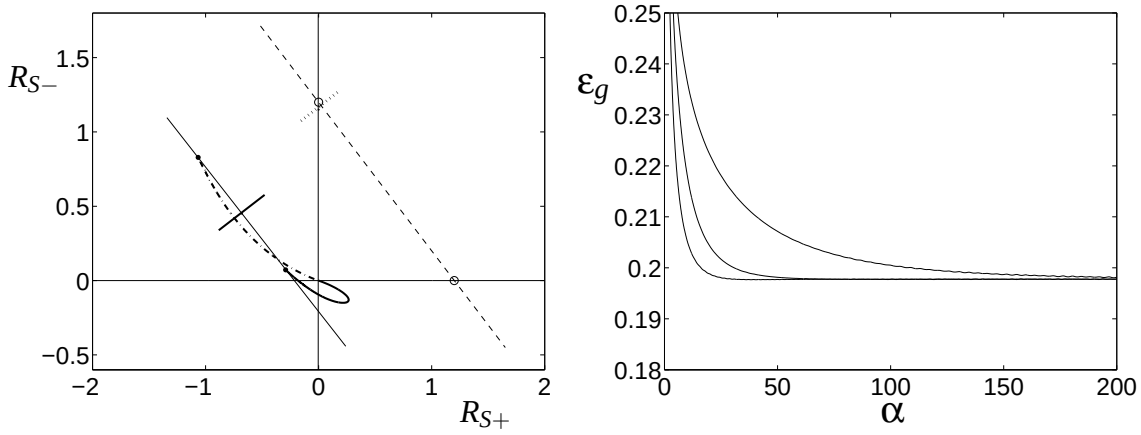

Figure 7: Learning from mistakes (LFM)

Left panel: Projected prototype trajectories for LFM with $\eta = 1$ and all other parameters as in Fig. 6 (left panel). The initial behavior for small α is similar to that of LVQ+/-. Asymptotically, both prototypes approach stationary positions (solid dots) which are not on the symmetry axis of the distribution (dashed line). Note that only the projections of $(w_+ - w_-)$ become parallel to $\lambda(B_+ - B_-)$ as $\alpha \to \infty$. While the actual $w_\pm(\alpha \to \infty)$ depend on η as discussed in the text, the asymptotic decision boundary (short solid line) does not.

Right panel: $\varepsilon_g(\alpha)$ for LFM with model parameters $\lambda = 3, p_+ = 0.8, v_+ = 4, v_- = 9$ and learning rates $\eta = 2.0, 1.0, 0.5$ (from left to right). All learning curves approach the same η-independent asymptotic value ε_g^{lfm}. Note the offset on the ε_g- axis.

Provided the decision boundary becomes similar to the Bayes optimal one in the course of learning, examples ξ from this region would satisfy $p_+ P(\xi|+1) \approx p_- P(\xi|-1)$, then. In effect, the system would be trained from balanced data, which significantly slows down the divergence, see the above considerations for LVQ+/- with $p_+ = p_-$.

In principle, these strategies can also be studied within our framework by considering appropriate modifications of the modulation function, Eq. (5). Preliminary results indicate, however, that both ideas improve convergence and generalization ability of LVQ+/- only to a very limited extent. In addition, one or several parameters, for example, the cut-off distance d_{max} or the *window size*, have to be carefully tuned to achieve the desired effect.

Moreover, both variants are generically outperformed by a conceptually simpler *early stopping* strategy. The idea is to stop the learning process as soon as the divergent behavior starts to increase the generalization error. This does not require the fine-tuning of additional parameters. However, in any practical situation, one would have to monitor ε_g in a test set of data which is not used for training.

Generic learning curves of LVQ+/-, see Figure 5, display a pronounced minimum before they approach the trivial asymptotic value $\varepsilon_g = \min\{p_+, p_-\}$. We are interested in the best generalization error ε_g^{stop} that could be achieved, in principle, by an idealized early stopping method. In contrast to $(\alpha \to \infty)$-asymptotic properties, the dynamics for intermediate α will depend on initialization,

as discussed below. This concerns also the depth of the minimum, which can occur at rather small α. For the cluster geometry considered here and initialization as given in Eq. (13), we find that the lowest values of ε_g are indeed achieved for $\widehat{Q} = 0$. In this setting the effect of the first examples is a very fast alignment of $(w_+ - w_-)$ with the symmetry axis of the model density (1). Hence, we first consider prototypes which are initialized precisely in the origin $w_+(0) = w_-(0) = 0$ and without offset from the plane spanned by B_+ and B_-. We observe furthermore that the respective value of ε_g in the minimum decreases with decreasing η. For the comparison with other algorithms we will therefore resort to the simultaneous limit $\eta \to 0, \alpha \to \infty$ as for LVQ1.

In our model, it is straightforward to obtain precisely the minimum of the limiting $\varepsilon_g(\widetilde{\alpha})$ from the analytic solution, Eq. (37). The result ε_g^{stop} is displayed as a function of p_+ in Figure 4 for two different choices of v_+ and v_-. In the case of equal variances $v_+ = v_-$ we find that LVQ+/- with early stopping is relatively close to the best possible ε_g^{bld}. However, it is outperformed by the asymptotic result of LVQ1 for all p_+. Note that both algorithms yield the optimal classification scheme in the singular case $p_+ = p_- = 1/2$ and $v_+ = v_-$.

For data with $v_+ \neq v_-$, we find that $\varepsilon_g^{stop} > \varepsilon_g^{lvq1}$ for small and large values of p_+, see Figure 4 (right panel) for an example. In the case of balanced priors, the learning curve of LVQ+/- does not display a minimum but approaches the optimal value ε_g^{bld} quite rapidly. For cluster weights in an interval around $p_+ = 1/2$, the result of the early stopping procedure is superior to that of LVQ1.

It is important to note that our analysis refers to an idealized early stopping procedure based on perfect knowledge of the current ε_g. On the contrary, the basic LVQ1 gives close to optimal performance independent of initial conditions, without adjustment of parameters and without explicit estimation of the generalization error.

Our results show that LVQ+/- with favorable initialization outperforms other algorithms with respect to behavior for small α, that is, with respect to learning from small training sets. Note, for instance, that the minima in the learning curve $\varepsilon_g(\alpha)$ as displayed in Fig. 5 can occur at very small α, already. While LVQ1, for example, achieves better $\alpha \to \infty$ asymptotic generalization, a comparison of the learning curves shows that it is often inferior for small and intermediate values of α. A systematic comparison is difficult, however, since the effect will depend strongly on the considered initialization. Nevertheless, the use of LVQ+/- type updates in the early stages of training or for limited availability of example data appears promising.

The crucial influence of the initialization is illustrated in Fig. 6. As an example, we consider the initialization of prototypes with an offset from the origin and, more importantly, from the plane spanned by the cluster centers. Again, learning curves display a minimum in the generalization error. However, the obtained best value of ε_g^{stop} is far from optimal, as displayed in Fig. 6 (right panel). Asymptotic properties of the considered algorithms are independent of initial settings and, hence, represent generic behavior in the frame of our model situation. On the contrary, the above discussed results for *early stopping* are highly sensitive with respect to initialization and demonstrate, at best, the potential usefulness of the strategy.

4.3 Learning from Mistakes

In the following, results are presented for the LFM training algorithm with and without a selective window.

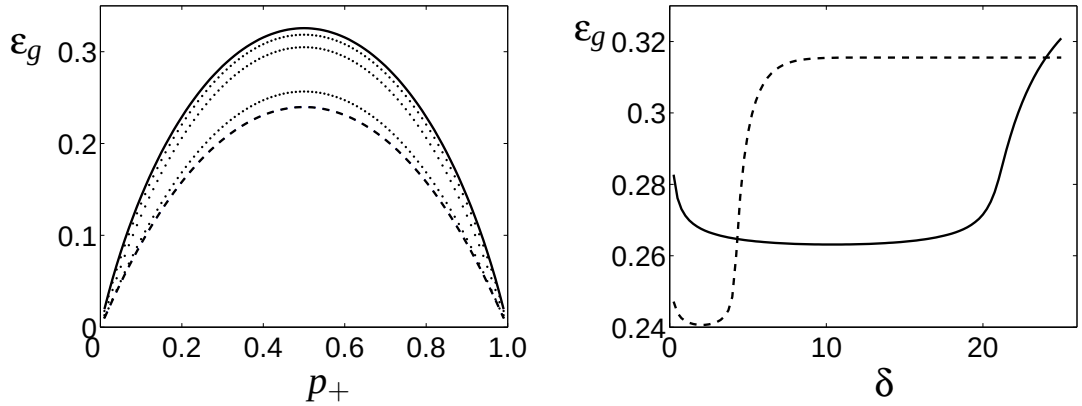

Figure 8: Learning from mistakes with window (LFM-W)
Left panel: Learning from equal variance clusters with $v_+ = v_= - 1$, $\lambda = 1$, and learning rate $\eta = 1$. Achieved $(\alpha \to \infty)$ asymptotic generalization error ε_g^{lfm-w} as a function of p_+ for different rescaled window sizes; from top to bottom: $\delta \to \infty$ (solid line, LFM), $\delta = 6$, $\delta = 5$, and $\delta = 0.25$ (dotted lines). The lowest, dashed line marks the best possible ε_g^{bld}.
Right panel: Asymptotic generalization error ε_g^{lfm-w} as a function of the window size δ for prior weight $p_+ = 0.6$ and learning rate $\eta = 1$. The dashed line corresponds to data with $v_+ = v_- = 1$ and $\lambda = 1$, the solid line refers to $v_+ = v_- = 5$ and $\lambda = 2$. Already in symmetric settings, the location of the optimal window size depends strongly on the properties of the data. Note the offset on the ε_g-axis.

4.3.1 LFM

The basic idea of the LFM procedure, cf. Section 2.3 (c), is reminiscent of many prescriptions for perceptron training (e.g., Watkin et al., 1993; Engel and van den Broeck, 2001; Biehl and Caticha, 2003). An LVQ+/- type of update is performed only if the current configuration would misclassify the example. Numerical integration of the corresponding ODE, Eq. (36), shows that LFM does not display the divergent behavior of LVQ+/-.

In contrast to LVQ1 and LVQ+/-, the typical learning curves of LFM display a monotonic decrease of $\varepsilon_g(\alpha)$, see the right panel of Figure 7 for examples with three different learning rates. An important feature is that the $(\alpha \to \infty)$-asymptotic generalization error of LFM learning does not depend on η.

Figure 7 (left panel) shows example trajectories of the prototypes in the $B_\pm$-plane. The behavior for small α is similar to that of LVQ+/-: Prototypes move away from the origin in opposite directions, initially. However, the attraction to (or repulsion from) the cluster centers becomes less important in the course of learning. Emphasis is on correct classification of the data, the aspect of cluster representation in the sense of VQ is essentially absent.

Eventually, the projections of the prototypes assume positions along a line which is parallel to the symmetry axis $\lambda(B_+ - B_-)$. However, the violation of condition (16) indicates that w_+ and w_- themselves do not lie in the two-dimensional subspace spanned by the cluster centers. Here,

the stationarity of order parameters and generalization error does not imply convergence of the prototype vectors themselves. Even in the limit $\alpha \to \infty$, they will fluctuate within the subspace of $\mathbb{R}^N \times \mathbb{R}^N$ that is compatible with the asymptotic values of $\{R_{S\sigma}, Q_{ST}\}$.

Clearly, the precise trajectory and the final outcome of LFM depends on the initialization of prototypes. It will determine, for instance, the relative position of w_+ and w_- parallel to the connection of cluster centers. The actual asymptotic configuration of order parameters and prototypes depends furthermore on the learning rate η. However, all these configurations share the same generalization error. The learning rate merely controls the magnitude of the fluctuations orthogonal to $\{B_+, B_-\}$ and the asymptotic distance of prototypes from the decision boundary. This corresponds to the observation that the following combinations of order parameters become proportional to η in the limit $\alpha \to \infty$:

$$r_+ = R_{++} - R_{-+}, \;\; r_- = R_{--} - R_{+-}, \;\; q = Q_{++} - Q_{--}, \;\; \delta = \sqrt{Q_{++} - 2Q_{+-} + Q_{--}} \qquad (17)$$

which results in an η-independent asymptotic generalization error, Eqs. (14,15). This finding also implies that the angle between $(w_+ - w_-)$ and $(B_+ - B_-)$ which is given by $\arccos\left[(r_+ + r_-)/(\sqrt{2}\delta)\right]$ becomes independent of the learning rate for $\alpha \to \infty$.

The quantity δ in (17) measures the distance $|w_+ - w_-|$ and vanishes for $\eta \to 0$. Hence, the prototypes will coincide and the consideration of this limit together with $\alpha \to \infty$ is not useful in the case of LFM.

The mathematical analysis of the stationary state reveals another interesting feature: One can show that fixed point configurations of the system of ODE for LFM, Eq. (36), satisfy the necessary condition

$$p_+ \varepsilon_+ \;=\; p_- \varepsilon_-.$$

That is, the two contributions to the total ε_g, Eqs. (14,15), become equal for $\alpha \to \infty$. As a consequence, LFM updates will be based on balanced data, asymptotically, as they are restricted to misclassified examples.

While learning from mistakes appears to be a very plausible concept and cures the divergence of LVQ+/-, the resulting asymptotic generalization ability ε_g^{lfm} turns out rather poor. The chain lines in Figure 4 mark ε_g^{lfm} for two different model scenarios. Not only is LFM outperformed by the basic LVQ1 and LVQ+/- with early stopping, for small and large values of p_+ its generalization error can even exceed the trivial value $\min\{p_+, p_-\}$.

It is important to note that the above results apply only to the crisp (LFM) version of RSLVQ (Seo and Obermayer, 2003). It is very well possible that truly *soft* schemes display a much better generalization behavior in our model situation. In fact, it has been demonstrated that RSLVQ performs significantly better than the crisp LFM and other LVQ algorithms in practical situations (Seo and Obermayer, 2003).

4.3.2 LFM-W

Our analysis shows that the introduction of a window for the selection of data, as defined in Sec. 2.3 (d), bears the potential to improve the generalization behavior of LFM drastically. The mathematical treatment is to a large extent analogous to that of the unrestricted LFM procedure, see Appendix A for details.

Qualitatively, the typical learning curves $R_{S\tau}(\alpha), Q_{ST}(\alpha)$ of LFM-W and the corresponding projected trajectories also resemble those of unrestricted LFM. Note, however, that the stationary generalization error ε_g^{lfm-w} of LFM-W does depend on the learning rate η, in contrast to the results displayed in Figure 7 (right panel) for LFM without window.

A detailed discussion of LFM-W will be presented elsewhere. Here, we focus on the role of the window parameter δ. Figure 8 (right panel) displays the dependence of the $(\alpha \to \infty)$ asymptotic ε_g^{lfm-w} for constant learning rate $\eta = 1.0$ and example settings of the model parameters. For very large δ, the suboptimal results of unrestricted LFM are recovered and also for $\delta \to 0$ the performance deteriorates. Evidently, an optimal choice of δ exists which yields the best generalization behavior, given all other parameters. Note that the precise position of the optimum and the robustness with respect to variation of δ depend strongly on the properties of the data.

We restrict ourselves to demonstrating that the performance can improve drastically in comparison with unrestricted LFM. Figure 8 (left panel), shows the stationary generalization error as a function of the prior weight p_+ for several (fixed) window sizes in an example setting. Note that ε_g^{lfm-w} for properly chosen values of δ is significantly lower than that for unrestricted LFM, that is, $\delta \to \infty$.

The evaluation of the truly optimal generalization error in the frame of our model is beyond the scope of this publication. In a specific learning problem, that is, for a particular choice of $\lambda, v_\pm$ and $p_\pm$ in our model, the training algorithm is to be optimized with respect to the two-dimensional parameter space of η and δ. Such an optimization would be difficult to achieve in practical situations and require sophisticated validation techniques.

5. Summary and Conclusion

We have rigorously investigated several basic LVQ algorithms: the original LVQ1, LVQ+/- with and without early stopping, *learning from mistakes* (LFM), and LFM-W which includes an additional window rule. The analysis is performed by means of the theory of on-line learning in a simple though relevant model setting.

It can be seen that LVQ+/- usually displays a divergent behavior whereas LVQ1 and LFM procedures converge towards fixed positions of the order parameters. These findings correspond to observations in practical situations.

The respective convergence speed depends on the learning rate in all cases. The same holds for the quality of the resulting classifier for LVQ1, LVQ+/- with early stopping, and LFM-W. For LFM without a selective window, on the contrary, the generalization ability of the stationary setting is independent of the choice of the learning rate. It should be mentioned that the trajectories of the prototypes need not be the shortest towards the final position and, often, initial overshooting as in LVQ1 can be observed if the classes are not balanced.

Even more interesting than their dynamical behavior is the generalization ability achieved by the algorithms. LVQ1 turns out to yield surprisingly good results, not very far from the optimum achievable error for this class of problems.

The outcome of LVQ+/- with early stopping can be close to or even slightly better than the $\alpha \to \infty$ asymptotic performance of LVQ1. However, it depends strongly on initialization and the detailed properties of the data. LVQ+/- like strategies appear particularly promising for the initial phase of training or when only a very limited training set is available.

The robustness of LFM as a crisp version of RSLVQ is clearly demonstrated by the fact that its asymptotic behavior is not affected by the magnitude of the learning rate. However, quite unexpectedly, it shows rather poor generalization ability, which is even inferior to a trivial classification for highly unbalanced data. We demonstrate that a proper selection of examples close to the current decision boundary as in LFM-W can improve the generalization performance drastically. Presumably, similar improvement could be achieved by employing RSLVQ with a soft assignment as suggested by Seo and Obermayer (2003). In practical situations, both schemes would require the careful tuning of a parameter, that is, the *softness* or *window size,* in order to achieve good generalization. We will address the detailed analysis and optimization of LFM-W and RSLVQ in forthcoming studies.

A few of our results, for instance the instability of LVQ+/-, may appear plausible from elementary reasoning. Others are clearly far from obvious and demonstrate the usefulness of our systematic approach. We show, in particular, the good generalization ability of the original LVQ1 learning rule. It does not follow the gradient of a well-defined cost function but outperforms several alternative algorithms. The relatively poor performance of the highly intuitive LFM training constitutes another non-trivial insight obtained from our analysis.

Our model is currently restricted to two unimodal classes and two prototypes, a case in which the mathematical analysis is feasible but which is far away from typical settings in practice. Nevertheless, this investigation yields relevant insights into practical situations. One can expect that an algorithm which does not yield a good generalization ability in this idealized scenario is also inappropriate for more complex, practical applications. In this sense, the investigation provides a meaningful testing ground for the success of learning algorithms.

Frequently, at most two prototypes are updated at a given time step also in larger LVQ networks. Hence, such systems can be interpreted as a superposition of pairs of prototypes within classes and at class borders. Within classes, a simple vector quantization takes place, a scenario which has been investigated using the same framework (Biehl et al., 1997; Ghosh et al., 2004). At cluster borders, the situation investigated in this article is relevant. However, the role of further, complicating effects in the dynamics of the superposition of these simple subsystems has to be studied in forthcoming projects. Also, the explicit extension of the theoretical framework to multi-class, multi-prototype problems is feasible under simplifying assumptions. It is currently in the focus of our efforts.

Based on the formalism presented in this article, a variety of further research becomes possible. Naturally, the consideration of alternative LVQ schemes is necessary. Learning rules which also change the metric during training such as the one proposed by Hammer and Villmann (2002) seem particularly interesting. However, it is not obvious how highly nonlinear adaptation schemes or algorithms which single out specific data components can be treated within our theoretical framework.

The ultimate goal of this work is the design of robust and reliable LVQ algorithms which provide optimal generalization ability. One step into this direction would be the on-line optimization of algorithm parameters, for example, the learning rate, based on the observed behavior in the course of training. Note that this can be done systematically by means of an optimization of the learning curve with respect to the learning rate in our setting. Even more promising, one can formally optimize the actual update functions f_S of the learning rule with respect to the generalization ability gained per example. This should be possible along the lines of variational optimization as it has been applied in the training of perceptrons or multilayered neural networks (e.g., Saad, 1999; Engel and van den Broeck, 2001; Biehl and Caticha, 2003). An alternative could be the construction of on-line prescriptions within a Bayesian framework, as it has been employed in the context of

perceptron training (for further references consult: Saad, 1999). Even if practical implementations turn out to be difficult, these investigations would help to identify the most important features of successful algorithms. Hence, this line of research should strengthen the mathematical foundation of LVQ training.

Appendix A. The Theoretical Framework

Here we outline key steps of the calculations referred to in the text. Important aspects of the formalism were first used in the context of unsupervised vector quantization (Biehl et al., 1997). Some of the calculations presented here were recently detailed in a Technical Report (Ghosh et al., 2004).

Throughout this appendix indices $l,m,k,s,\sigma \in \{\pm 1\}$ (or $\pm$ for short) represent the class labels and cluster memberships. We also use the shorthand

$$\Theta_s = \Theta\left(d_{-s} - d_{+s}\right)$$

for the Heaviside function in LVQ1. We furthermore employ the notations

$$\widehat{\Theta}^{o}_{\sigma} = \Theta\left(d_{\sigma} - d_{-\sigma}\right) \quad \text{and} \quad \widehat{\Theta}^{\delta}_{\sigma} = \Theta\left(d_{\sigma} - d_{-\sigma} - \delta\right) \tag{18}$$

for the complementary Heaviside function in LFM and the modified update of LFM-W, respectively.

A.1 Statistics of the Projections

To a large extent, our analysis is based on the observation that the projections $h_{\pm} = \mathbf{w}_{\pm} \cdot \xi$ and $b_{\pm} = \mathbf{B}_{\pm} \cdot \xi$ are correlated Gaussian random quantities for a vector ξ drawn from one of the clusters contributing to the density (1). Where convenient, we will combine the projections into a four-dimensional vector denoted as $\underline{x} = (h_+, h_-, b_+, b_-)^T$.

We will assume implicitly that ξ is statistically independent from the considered weight vectors $w_{\pm}$. This is obviously the case in our on-line prescription where the novel example ξ^{μ} is uncorrelated with all previous data and hence with $w_{\pm}^{\mu-1}$. For the sake of readability we omit indices μ in the following.

The first and second conditional moments given in Eq. (11) are obtained from the following elementary considerations.

First Moments
Exploiting the above mentioned statistical independence we can show immediately that

$$\langle h_l \rangle_k = \langle w_l \cdot \xi \rangle_k = w_l \cdot \langle \xi \rangle_k = w_l \cdot (\lambda B_k) = \lambda R_{lk}. \tag{19}$$

Similarly we get for b_l:

$$\langle b_l \rangle_k = \langle B_l \cdot \xi \rangle_k = B_l \cdot \langle \xi \rangle_k = B_l \cdot (\lambda B_k) = \lambda \delta_{lk}, \tag{20}$$

where δ_{lk} is the Kronecker delta and we exploit that B_+ and B_- are orthonormal. Now the conditional means $\underline{\mu}_k = \langle \underline{x} \rangle_k$ can be written as

$$\underline{\mu}_{k=+1} = \lambda\left(R_{++}, R_{-+}, 1, 0\right)^T \qquad \text{and} \qquad \underline{\mu}_{k=-1} = \lambda\left(R_{+-}, R_{--}, 0, 1\right)^T. \tag{21}$$

Second Moments

In order to compute the conditional variance or covariance $\langle h_l h_m\rangle_k - \langle h_l\rangle_k\langle h_m\rangle_k$ we first consider the average

$$\begin{aligned}
\langle h_l\, h_m\rangle_k &= \langle (w_l\cdot\xi)(w_m\cdot\xi)\rangle_k = \left\langle \left(\sum_{i=1}^N (w_l)_i(\xi)_i\right)\left(\sum_{j=1}^N (w_m)_j(\xi)_j\right)\right\rangle_k \\
&= \left\langle \sum_{i=1}^N (w_l)_i(w_m)_i(\xi)_i(\xi)_i + \sum_{i=1}^N\sum_{j=1,j\neq i}^N (w_l)_i(w_m)_j(\xi)_i(\xi)_j\right\rangle_k \\
&= \sum_{i=1}^N (w_l)_i(w_m)_i\langle(\xi)_i(\xi)_i\rangle_k + \sum_{i=1}^N\sum_{j=1,j\neq i}^N (w_l)_i(w_m)_j\langle(\xi)_i(\xi)_j\rangle_k \\
&= \sum_{i=1}^N (w_l)_i(w_m)_i[v_k+\lambda^2(B_k)_i(B_k)_i] + \sum_{i=1}^N\sum_{j=1,j\neq i}^N (w_l)_i(w_m)_j\lambda^2(B_k)_i(B_k)_j \\
&= v_k\sum_{i=1}^N (w_l)_i(w_m)_i + \lambda^2\sum_{i=1}^N (w_l)_i(w_m)_i(B_k)_i(B_k)_i \\
&\quad +\lambda^2\sum_{i=1}^N\sum_{j=1,j\neq i}^N (w_l)_i(w_m)_j(B_k)_i(B_k)_j \\
&= v_k w_l\cdot w_m + \lambda^2(w_l\cdot B_k)(w_m\cdot B_k) \;= v_k Q_{lm} + \lambda^2 R_{lk}R_{mk}\,.
\end{aligned}$$

Here we have used that components of ξ from cluster k have variance v_k and are independent. This implies for all $i,j\in\{1,\ldots,N\}$:

$$\langle(\xi)_i(\xi)_i\rangle_k - \langle(\xi)_i\rangle_k\langle(\xi)_i\rangle_k = v_k \;\;\Rightarrow\;\; \langle(\xi)_i(\xi)_i\rangle_k = v_k + \langle(\xi)_i\rangle_k\langle(\xi)_i\rangle_k\,,$$

and $\langle(\xi)_i(\xi)_j\rangle_k = \langle(\xi)_i\rangle_k\langle(\xi)_j\rangle_k$ for $i\neq j$.

Finally, we obtain the conditional second moment

$$\langle h_l\, h_m\rangle_k - \langle h_l\rangle_k\langle h_m\rangle_k = v_k Q_{lm} + \lambda^2 R_{lk}R_{mk} - \lambda^2 R_{lk}R_{mk} = v_k Q_{lm}\,.$$

In an analogous way we get

$$\langle b_l b_m\rangle_k - \langle b_l\rangle_k\langle b_m\rangle_k = v_k\delta_{lm} \quad\text{and}\quad \langle h_l b_m\rangle_k - \langle h_l\rangle_k\langle b_m\rangle_k = v_k R_{lm}\,.$$

The above results are summarized in Eq. (11). The conditional covariance matrix of $\underline{x}$ can be expressed explicitly in terms of the order parameters as follows:

$$C_k = v_k\begin{pmatrix} Q_{++} & Q_{+-} & R_{++} & R_{+-} \\ Q_{+-} & Q_{--} & R_{-+} & R_{--} \\ R_{++} & R_{-+} & 1 & 0 \\ R_{+-} & R_{--} & 0 & 1 \end{pmatrix}.$$

The conditional density of $\underline{x}$ for data from class k is a Gaussian $N(\underline{\mu}_k, C_k)$ where $\underline{\mu}_k$ is the conditional mean vector, Eq. (21), and C_k is given above.

A.2 Form of the Differential Equations

Here, the differential equations for $\{R_{s\sigma}, Q_{st}\}$ are given for LVQ1, LVQ+/-, LFM, and LFM-W, before Gaussian averages are performed.

A.2.1 LVQ1

In the case of the LVQ1 algorithm the generic form of the coupled system of ODE (Eq. (12)) yields the following system:

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta\left(l\Big(\langle\sigma b_m\Theta_l\rangle - \langle\sigma\Theta_l\rangle R_{lm}\Big)\right),\\
\frac{dQ_{lm}}{d\alpha} &= \eta\left(l\Big(\langle\sigma h_m\Theta_l\rangle - \langle\sigma\Theta_l\rangle Q_{lm}\Big) + m\Big(\langle\sigma h_l\Theta_m\rangle - \langle\sigma\Theta_m\rangle Q_{lm}\Big)\right.\\
&\qquad \left. + \eta\,\delta_{lm}\sum_{\sigma=\pm1} v_\sigma p_\sigma\langle\Theta_l\rangle_\sigma\right).
\end{aligned} \tag{22}$$

A.2.2 LVQ+/-

In the case of LVQ+/- Eq. (12) results in the following system of coupled ODE:

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta\, l\left(\langle\sigma b_m\rangle - \langle\sigma\rangle R_{lm}\right),\\
\frac{dQ_{lm}}{d\alpha} &= \eta\left(l\langle\sigma h_m\rangle - l\langle\sigma\rangle Q_{lm} + m\langle\sigma h_l\rangle - m\langle\sigma\rangle Q_{lm} + \eta\, l\, m\sum_{\sigma=\pm1} p_\sigma v_\sigma\right).
\end{aligned} \tag{23}$$

A.2.3 LFM

With the modulation function for LFM, Eq. (12), and using $\widehat{\Theta}^o_\sigma$ from Eq. (18) we obtain

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta\, l\left(\langle\sigma b_m\widehat{\Theta}^o_\sigma\rangle - \langle\sigma\widehat{\Theta}^o_\sigma\rangle R_{lm}\right), \qquad (24)\\
\frac{dQ_{lm}}{d\alpha} &= \eta\left(l\langle\sigma h_m\widehat{\Theta}^o_\sigma\rangle - (l+m)\langle\sigma\widehat{\Theta}^o_\sigma\rangle Q_{lm} + m\langle\sigma h_l\widehat{\Theta}^o_\sigma\rangle + l\,m\eta\sum_{\sigma=\pm1} p_\sigma v_\sigma\langle\widehat{\Theta}^o_\sigma\rangle_\sigma\right).
\end{aligned}$$

A.2.4 LFM-W

For the modulation function of the LFM-W algorithm in Eq. (12) we obtain

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta\, l\left(\langle\sigma b_m(\widehat{\Theta}^o_\sigma - \widehat{\Theta}^\delta_\sigma)\rangle - \langle\sigma(\widehat{\Theta}^o_\sigma - \widehat{\Theta}^\delta_\sigma)\rangle R_{lm}\right), \qquad (25)\\
\frac{dQ_{lm}}{d\alpha} &= \eta\left(l\langle\sigma h_m(\widehat{\Theta}^o_\sigma - \widehat{\Theta}^\delta_\sigma)\rangle - (l+m)\langle\sigma(\widehat{\Theta}^o_\sigma - \widehat{\Theta}^\delta_\sigma)\rangle Q_{lm} + m\langle\sigma h_l(\widehat{\Theta}^o_\sigma - \widehat{\Theta}^\delta_\sigma)\rangle\right.
\end{aligned}$$

$$+ l\,m\eta \sum_{\sigma=\pm 1} p_\sigma v_\sigma \langle (\widehat{\Theta}_\sigma^o - \widehat{\Theta}_\sigma^\delta) \rangle_\sigma \Bigg),$$

where $\widehat{\Theta}_\sigma^o$ and $\widehat{\Theta}_\sigma^\delta$ are defined in Eq. (18). Note that for $\delta \to \infty$ we recover Eqs. (24) for LFM as $\lim_{\delta\to\infty} \widehat{\Theta}^\delta = 0$.

A.3 Gaussian Averages

In order to obtain the actual ODE for a given modulation function, averages over the joint density $P(h_+, h_-, b_+, b_-)$ are performed for LVQ1, LVQ+/-, and LFM.

A.3.1 LVQ+/-

The elementary averages in Eq. (23) are directly obtained from Eqs. (19,20) and read:

$$\langle \sigma b_m \rangle = \sum_{\sigma=\pm 1} \sigma p_\sigma \lambda \delta_{m,\sigma}, \qquad \langle \sigma h_m \rangle = \sum_{\sigma=\pm 1} \sigma p_\sigma \lambda R_{m,\sigma}, \qquad \langle \sigma \rangle = \sum_{\sigma=\pm 1} \sigma p_\sigma. \tag{26}$$

A.3.2 LVQ1

In the systems of ODE presented in Eq. (22) we encounter Heaviside functions of the following generic form:

$$\Theta_s = \Theta(\underline{\alpha}_s . \underline{x} - \beta_s)$$

which gives $\Theta_s = \Theta\left(d_{-s} - d_{+s}\right) = \Theta(\underline{\alpha}_s . \underline{x} - \beta_s)$ with

$$\underline{\alpha}_s = (+2s, -2s, 0, 0)^T \text{ and } \beta_s = \left(Q_{+s+s} - Q_{-s-s}\right), \tag{27}$$

Performing the averages in Eqs. (22) involves conditional means of the form

$$\langle (\underline{x})_n \Theta_s \rangle_k \quad \text{and} \quad \langle \Theta_s \rangle_k$$

where $(\underline{x})_n$ is the n^{th} component of $\underline{x} = (h_{+1}, h_{-1}, b_{+1}, b_{-1})$. We first address the term

$$\langle (\underline{x})_n \Theta_s \rangle_k = \frac{(2\pi)^{-2}}{(det(C_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} (\underline{x})_n \Theta\left(\underline{\alpha}_s \cdot \underline{x} - \beta_s\right) \exp\left(-\frac{1}{2}\left(\underline{x} - \underline{\mu}_k\right)^T C_k^{-1}\left(\underline{x} - \underline{\mu}_k\right)\right) d\underline{x}$$

$$= \frac{(2\pi)^{-2}}{(det(C_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} \left(\underline{x}' + \underline{\mu}_k\right)_n \Theta\left(\underline{\alpha}_s \cdot \underline{x}' + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s\right) \exp\left(-\frac{1}{2}\underline{x}'^T C_k^{-1} \underline{x}'\right) d\underline{x}'$$

with the substitution $\underline{x}' = \underline{x} - \underline{\mu}_k$.

Let $\underline{x}' = C_k^{\frac{1}{2}} \underline{y}$, where $C_k^{\frac{1}{2}}$ is defined in the following way: $C_k = C_k^{\frac{1}{2}} C_k^{\frac{1}{2}}$. Since C_k is a covariance matrix, it is positive semidefinite and $C_k^{\frac{1}{2}}$ exists. Hence we have $d\underline{x}' = det(C_k^{\frac{1}{2}}) d\underline{y} = (det(C_k))^{\frac{1}{2}} d\underline{y}$

and
$\langle (\underline{x})_n \Theta_s \rangle_k =$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (C_k^{\frac{1}{2}} \underline{y})_n \Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \exp\left(-\frac{1}{2} \underline{y}^2 \right) d\underline{y} + (\underline{\mu}_k)_n \langle \Theta_s \rangle_k \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} \sum_{j=1}^{4} \left((C_k^{\frac{1}{2}})_{nj} (\underline{y})_j \right) \Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \exp\left(-\frac{1}{2} \underline{y}^2 \right) d\underline{y} + (\underline{\mu}_k)_n \langle \Theta_s \rangle_k \\
&= I + (\underline{\mu}_k)_n \langle \Theta_s \rangle_k \qquad \text{(introducing the abbreviation } I\text{).} \qquad (28)
\end{aligned}
$$

Now consider the integrals contributing to I:

$$
I_j = \int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} (\underline{y})_j \Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right) d(\underline{y})_j .
$$

We can perform an integration by parts, $\int u dv = uv - \int v du$, with
$u = \Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right), \qquad v = (C_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right)$
$du = \frac{\partial}{\partial (\underline{y})_j} \Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) d(\underline{y})_j$
$dv = (-)(C_k^{\frac{1}{2}})_{nj} (\underline{y})_j \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right) d(\underline{y})_j$, and obtain

$$
\begin{aligned}
I_j &= \underbrace{- \left[\Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) (C_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right) \right]_{-\infty}^{\infty}}_{0} \\
&\quad + \left[\int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} \frac{\partial}{\partial (\underline{y})_j} \left(\Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \right) \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right) d(\underline{y})_j \right] \\
&= \int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} \frac{\partial}{\partial (\underline{y})_j} \left(\Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \right) \exp\left(-\frac{1}{2} (\underline{y})_j^2 \right) d(\underline{y})_j .
\end{aligned}
$$

In total we get

$$
\begin{aligned}
I &= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} (C_k^{\frac{1}{2}})_{nj} \int_{\mathbb{R}^4} \frac{\partial}{\partial (\underline{y})_j} \left(\Theta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \right) \exp\left(-\frac{1}{2} \underline{y}^2 \right) d\underline{y} \\
&= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left((C_k^{\frac{1}{2}})_{nj} \sum_{i=1}^{4} (\underline{\alpha}_s)_i (C_k^{\frac{1}{2}})_{i,j} \right) \int_{\mathbb{R}^4} \delta \left(\underline{\alpha}_s \dot{C}_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \exp\left(-\frac{1}{2} \underline{y}^2 \right) d\underline{y}. \\
&= \frac{1}{(2\pi)^2} (C_k \underline{\alpha}_s)_n \int_{\mathbb{R}^4} \left(\delta \left(\underline{\alpha}_s \cdot C_k^{\frac{1}{2}} \underline{y} + \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s \right) \right) \exp\left(-\frac{1}{2} \underline{y}^2 \right) d\underline{y}.
\end{aligned}
$$

In the last step we have used

$$\frac{\partial}{\partial(\underline{y})_j}\Theta\left(\underline{\alpha}_s\cdot C_k^{\frac{1}{2}}\underline{y}+\underline{\alpha}_s\cdot\underline{\mu}_k-\beta_s\right)=\sum_{i=1}^{4}(\underline{\alpha}_s)_i(C_k^{\frac{1}{2}})_{i,j}\delta(\underline{\alpha}_s\cdot C_k^{\frac{1}{2}}\underline{y}+\underline{\alpha}_s\cdot\underline{\mu}_k-\beta_s)$$

with the Dirac delta-function $\delta(.)$.

Now, note that $\exp\left[-\frac{1}{2}\underline{y}^2\right]d\underline{y}$ is a measure which is invariant under rotation of the coordinate axes. We rotate the system in such a way that one of the axes, say $\widetilde{y}$, is aligned with the vector $C_k^{\frac{1}{2}}\underline{\alpha}_s$. The remaining three coordinates can be integrated over and we get

$$I=\frac{1}{\sqrt{2\pi}}(C_k\underline{\alpha}_s)_n\int_{\mathbb{R}}\delta\left(\|C_k^{\frac{1}{2}}\underline{\alpha}_s\|\widetilde{y}+\underline{\alpha}_s\cdot\underline{\mu}_k-\beta_s\right)\exp\left[-\frac{1}{2}\widetilde{y}^2\right]d\widetilde{y}.$$

We define

$$\widetilde{\alpha}_{sk}=\|C_k^{\frac{1}{2}}\underline{\alpha}_s\|=\sqrt{\underline{\alpha}_s\cdot C_k\underline{\alpha}_s}\quad\text{and}\quad\widetilde{\beta}_{sk}=\underline{\alpha}_s\cdot\underline{\mu}_k-\beta_s \tag{29}$$

and obtain

$$\begin{aligned}
I &= \frac{1}{\sqrt{2\pi}}(C_k\underline{\alpha}_s)_n\int_{\mathbb{R}}\delta\left(\widetilde{\alpha}_{sk}\widetilde{y}+\widetilde{\beta}_{sk}\right)\exp\left[-\frac{1}{2}\widetilde{y}^2\right]d\widetilde{y}\\
&= \frac{(C_k\underline{\alpha}_s)_n}{\sqrt{2\pi}\widetilde{\alpha}_{sk}}\int_{\mathbb{R}}\delta\left(z+\widetilde{\beta}_{sk}\right)\exp\left[-\frac{1}{2}\left(\frac{z}{\widetilde{\alpha}_{sk}}\right)^2\right]dz\quad(\text{with}\quad z=\alpha_{sk}\widetilde{y})\\
&= \frac{(C_k\underline{\alpha}_s)_n}{\sqrt{2\pi}\widetilde{\alpha}_{sk}}\exp\left[-\frac{1}{2}\left(\frac{\widetilde{\beta}_{sk}}{\widetilde{\alpha}_{sk}}\right)^2\right].
\end{aligned} \tag{30}$$

Now we compute the remaining average in (28) in an analogous way and get

$$\begin{aligned}
\langle\Theta_s\rangle_k &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\Theta\left(\widetilde{\alpha}_{sk}\widetilde{y}+\widetilde{\beta}_{sk}\right)\exp\left[-\frac{1}{2}\widetilde{y}^2\right]d\widetilde{y}\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\widetilde{\beta}_{sk}}{\widetilde{\alpha}_{sk}}}\exp\left[-\frac{1}{2}\widetilde{y}^2\right]d\widetilde{y}=\Phi\left(\frac{\widetilde{\beta}_{sk}}{\widetilde{\alpha}_{sk}}\right)\quad\text{with}\quad\Phi(z)=\int_{-\infty}^{z}dx\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.
\end{aligned} \tag{31}$$

Finally we obtain the required average using (30) and (31) as follows:

$$\langle(\underline{x})_n\Theta_s\rangle_k = \frac{(C_k\underline{\alpha}_s)_n}{\sqrt{2\pi}\widetilde{\alpha}_{sk}}\exp\left[-\frac{1}{2}\left(\frac{\widetilde{\beta}_{sk}}{\widetilde{\alpha}_{sk}}\right)^2\right]+(\underline{\mu}_k)_n\Phi\left(\frac{\widetilde{\beta}_{sk}}{\widetilde{\alpha}_{sk}}\right). \tag{32}$$

The quantities $\widetilde{\alpha}_{sk}$ and $\widetilde{\beta}_{sk}$ are defined through Eq. (29) and (27) for LVQ1.

A.3.3 LFM AND LFM-W

In Eqs. (24,25) we have to evaluate conditional means of the form

$$\langle(\underline{x})_n\widehat{\Theta}_s^{\delta}\rangle_k,\ \langle\widehat{\Theta}_s^{\delta}\rangle_k\ \text{with the special cases}\ \langle(\underline{x})_n\widehat{\Theta}_s^{o}\rangle_k,\ \langle\widehat{\Theta}_s^{o}\rangle_k.$$

Here we use the notation

$$\begin{aligned}\widehat{\Theta}_\sigma^\delta &= \Theta(d_\sigma - d_{-\sigma} - \delta) = \Theta(\widehat{\underline{\alpha}}_\sigma.\underline{x} - \widehat{\beta}_\sigma^\delta)\\ \widehat{\Theta}_\sigma^o &= \Theta(d_\sigma - d_{-\sigma}) \qquad = \Theta(\widehat{\underline{\alpha}}_\sigma.\underline{x} - \widehat{\beta}_\sigma^o)\end{aligned}$$

with $\widehat{\underline{\alpha}}_\sigma = (-2\sigma, +2\sigma, 0, 0)^T$, $\widehat{\beta}_\sigma^\delta = -(Q_{\sigma\sigma} - Q_{-\sigma-\sigma} - \delta)$, and $\widehat{\beta}_\sigma^o = -(Q_{\sigma\sigma} - Q_{-\sigma-\sigma})$.

The mathematical structure is completely analogous to the case of LVQ1 and we obtain the results

$$\langle \widehat{\Theta}_s^\delta \rangle_k = \Phi\left(\frac{\widetilde{\beta}_{sk}^\delta}{\widehat{\alpha}_{sk}}\right) \text{ and } \langle (\underline{x})_n \widehat{\Theta}_s^\delta \rangle_k = \frac{(C_k \widehat{\underline{\alpha}}_s)_n}{\sqrt{2\pi}\widehat{\alpha}_{sk}} \exp\left[-\frac{1}{2}\Big(\frac{\widetilde{\beta}_{sk}^\delta}{\widehat{\alpha}_{sk}}\Big)^2\right] + (\underline{\mu}_k)_n \Phi\left(\frac{\widetilde{\beta}_{sk}^\delta}{\widehat{\alpha}_{sk}}\right), \qquad (33)$$

as well as the corresponding special cases with $\delta = 0$. Here, we have introduced

$$\widetilde{\beta}_{sk}^\delta = \widehat{\underline{\alpha}}_s \cdot \underline{\mu}_k - \widehat{\beta}_s^\delta, \;\; \widetilde{\beta}_{sk}^o = \widehat{\underline{\alpha}}_s \cdot \underline{\mu}_k - \widehat{\beta}_s^o, \text{ and } \widehat{\alpha}_{sk} = \sqrt{\widehat{\underline{\alpha}}_s \cdot C_k \widehat{\underline{\alpha}}_s}.$$

A.4 Final Form of the Differential Equations

The full form of the ODE for LVQ1, LVQ+/-, and LFM is obtained after inserting the averages given in the previous section.

A.4.1 LVQ1

For the LVQ1 algorithm, using (31) and (32), the system or ODE reads:

$$\begin{aligned}\frac{dR_{lm}}{d\alpha} &= \eta\Bigg[l\Big(\sum_{\sigma=\pm1}\sigma p_\sigma\Bigg[\frac{(C\underline{\alpha}_l)_{n_{bm}}}{\sqrt{2\pi}\widetilde{\alpha}_{l\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)^2\Big]\\ &\quad +(\underline{\mu}_\sigma)_{n_{bm}}\Phi\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)\Bigg] - \sum_{\sigma=\pm1}\sigma p_\sigma \Phi\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)R_{lm}\Big)\Bigg],\\ \frac{dQ_{lm}}{d\alpha} &= \eta\Bigg(l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\alpha}_l)_{n_{hm}}}{\sqrt{2\pi}\widetilde{\alpha}_{l\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)^2\Big] + (\underline{\mu}_\sigma)_{n_{hm}}\Phi\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)\Big]\\ &\quad -l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big)Q_{lm} + m\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\alpha}_l)_{n_{hl}}}{\sqrt{2\pi}\widetilde{\alpha}_{m\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}_{m\sigma}}{\widetilde{\alpha}_{m\sigma}}\Big)^2\Big]\\ &\quad +(\underline{\mu}_\sigma)_{n_{hl}}\Phi\Big(\frac{\widetilde{\beta}_{m\sigma}}{\widetilde{\alpha}_{m\sigma}}\Big)\Big] - m\sum_{\sigma=\pm1}\sigma p_\sigma\Phi\Big(\frac{\widetilde{\beta}_{m\sigma}}{\widetilde{\alpha}_{m\sigma}}\Big)Q_{lm}\Bigg)\\ &\quad +\delta_{lm}\eta^2\sum_{\sigma=\pm1}\sigma v_\sigma p_\sigma \Phi\Big(\frac{\widetilde{\beta}_{l\sigma}}{\widetilde{\alpha}_{l\sigma}}\Big). \qquad (34)\end{aligned}$$

Here, we use the previously defined abbreviations, Eq. (29),

$$\widetilde{\alpha}_{sk} = \sqrt{\underline{\alpha}_s \cdot C_k \underline{\alpha}_s}, \qquad \widetilde{\beta}_{sk} = \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s$$

with $\underline{\alpha}_s = (+2s, -2s, 0, 0)^T$ and $\beta_s = (Q_{+s+s} - Q_{-s-s})$.

Furthermore, $n_{hm} = \begin{cases} 1 & \text{if } m = 1 \\ 2 & \text{if } m = -1 \end{cases}$ and $n_{bm} = \begin{cases} 3 & \text{if } m = 1 \\ 4 & \text{if } m = -1 \end{cases}$.

A.4.2 LVQ+/-

Using the averages computed in (26) we get the final form of the system of ODE for the LVQ+/- algorithm as follows:

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta l \left(\sum_{\sigma=\pm 1} \sigma p_\sigma \lambda \delta_{m,\sigma} - \sum_{\sigma=\pm 1} \sigma p_\sigma R_{lm} \right), \\
\frac{dQ_{lm}}{d\alpha} &= \eta \Bigg(l \sum_{\sigma=\pm 1} \sigma p_\sigma \lambda R_{m,\sigma} - l \sum_{\sigma=\pm 1} \sigma p_\sigma Q_{lm} + m \sum_{\sigma=\pm 1} \sigma p_\sigma \lambda R_{l,\sigma} \\
&\quad - m \sum_{\sigma=\pm 1} \sigma p_\sigma Q_{lm} + \eta l m \sum_{\sigma=\pm 1} p_\sigma v_\sigma \Bigg).
\end{aligned} \tag{35}$$

A.4.3 LFM

The final form of the system of ODE reads with Eq. (33)

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \eta l \Bigg(\sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\frac{(C\widehat{\underline{\alpha}}_\sigma)_{n_{bm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}} \exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2 \Big] \\
&\quad + (\underline{\mu}_\sigma)_{n_{bm}} \Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] - \sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] R_{lm} \Bigg), \\
\frac{dQ_{lm}}{d\alpha} &= \eta \Bigg(l \sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\frac{(C\widehat{\underline{\alpha}}_\sigma)_{n_{hm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}} \exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2 \Big] + (\underline{\mu}_\sigma)_{n_{hm}} \Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] \\
&\quad - l \sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] Q_{lm} + m \sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\frac{(C\widehat{\underline{\alpha}}_\sigma)_{n_{hl}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}} \exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2 \Big] \\
&\quad + (\underline{\mu}_\sigma)_{n_{hl}} \Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] - m \sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Big] Q_{lm} + l m \eta \sum_{\sigma=\pm 1} v_\sigma p_\sigma \Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big) \Bigg).
\end{aligned} \tag{36}$$

Here we have to insert $\widehat{\alpha}_{sk} = \sqrt{\widehat{\underline{\alpha}}_s \cdot C_k \widehat{\underline{\alpha}}_s}$, $\quad \widetilde{\beta}^o_{sk} = \widehat{\underline{\alpha}}_s \cdot \underline{\mu}_k - \widehat{\beta}^o_s$

with $\quad \widehat{\underline{\alpha}}_\sigma = (-2\sigma, +2\sigma, 0, 0)^T$ and $\widehat{\beta}^o_\sigma = -(Q_{+\sigma+\sigma} - Q_{-\sigma-\sigma})$.

Also, $\quad n_{hm} = \begin{cases} 1 & \text{if } m = 1 \\ 2 & \text{if } m = -1 \end{cases}$ and $n_{bm} = \begin{cases} 3 & \text{if } m = 1 \\ 4 & \text{if } m = -1 \end{cases}$.

A.4.4 LFM-W

The system of ODE for LFM-W, using Eq. (33), is given by

$$\frac{dR_{lm}}{d\alpha} = \eta l \Bigg(\sum_{\sigma=\pm 1} \sigma p_\sigma \Big[\frac{(C\widehat{\underline{\alpha}}_\sigma)_{n_{bm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}} \exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2 \Big]$$

$$
\begin{aligned}
&+(\underline{\mu}_\sigma)_{n_{bm}}\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big] - \sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]R_{lm}\Bigg)\\
&-\eta\, l\Bigg(\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\widehat{\alpha}}_\sigma)_{n_{bm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2\Big]\\
&+(\underline{\mu}_\sigma)_{n_{bm}}\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big] - \sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]R_{lm}\Bigg),\\
\frac{dQ_{lm}}{d\alpha} \;=\;& \eta\Bigg(l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\widehat{\alpha}}_\sigma)_{n_{hm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2\Big] + (\underline{\mu}_\sigma)_{n_{hm}}\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]\\
&-l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]Q_{lm} + m\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\widehat{\alpha}}_\sigma)_{n_{hl}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2\Big]\\
&+(\underline{\mu}_\sigma)_{n_{hl}}\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big] - m\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]Q_{lm} + l\,m\,\eta\sum_{\sigma=\pm1}v_\sigma p_\sigma\Phi\Big(\frac{\widetilde{\beta}^o_{\sigma\sigma}}{\widetilde{\alpha}_{\sigma\sigma}}\Big)\Bigg)\\
&-\eta\Bigg(l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\widehat{\alpha}}_\sigma)_{n_{hm}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2\Big] + (\underline{\mu}_\sigma)_{n_{hm}}\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]\\
&-l\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]Q_{lm} + m\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\frac{(C\underline{\widehat{\alpha}}_\sigma)_{n_{hl}}}{\sqrt{2\pi}\widehat{\alpha}_{\sigma\sigma}}\exp\Big[-\frac{1}{2}\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)^2\Big]\\
&+(\underline{\mu}_\sigma)_{n_{hl}}\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big] - m\sum_{\sigma=\pm1}\sigma p_\sigma\Big[\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Big]Q_{lm} + l\,m\,\eta\sum_{\sigma=\pm1}v_\sigma p_\sigma\Phi\Big(\frac{\widetilde{\beta}^\delta_{\sigma\sigma}}{\widehat{\alpha}_{\sigma\sigma}}\Big)\Bigg).
\end{aligned}
$$

Here again, we have to insert

$$\widehat{\alpha}_{sk} = \sqrt{\underline{\widehat{\alpha}}_s\cdot C_k\,\underline{\alpha}_s},\quad \widetilde{\beta}^\delta_{sk} = \underline{\widehat{\alpha}}_s\cdot\underline{\mu}_k - \widehat{\beta}^\delta_s,\quad \widetilde{\beta}^o_{sk} = \underline{\widehat{\alpha}}_s\cdot\underline{\mu}_k - \widehat{\beta}^o_s$$

with $\underline{\widehat{\alpha}}_\sigma = (-2\sigma,+2\sigma,0,0)^T, \widehat{\beta}^\delta_\sigma = -(Q_{+\sigma+\sigma} - Q_{-\sigma-\sigma} - \delta)$, and $\widehat{\beta}^o_\sigma = -(Q_{+\sigma+\sigma} - Q_{-\sigma-\sigma})$.

As above, $\quad n_{hm} = \begin{cases} 1 & \text{if } m = 1 \\ 2 & \text{if } m = -1 \end{cases}$ and $n_{bm} = \begin{cases} 3 & \text{if } m = 1 \\ 4 & \text{if } m = -1 \end{cases}$.

A.5 Analytical Results for LVQ+/-

The system of ODE (35) can be integrated analytically and the solutions are presented below for initialization of prototypes in the origin, that is, $R_{lm}(0) = Q_{lm}(0) = 0$.
For convenience, we use the parameterization

$$p_+ = \frac{1+\widehat{p}}{2}, \qquad p_- = \frac{1-\widehat{p}}{2}, \quad \text{with the } \textit{bias}\ \ \widehat{p} = (2p_+ - 1) \in [-1,1].$$

In the generic case of unequal priors $\widehat{p} \neq 0$, one obtains

$$
\begin{aligned}
R_{++}(\alpha) &= -\frac{1}{2\widehat{p}}\lambda(1+\widehat{p})\,(e^{-\alpha\eta\widehat{p}} - 1), \qquad R_{+-}(\alpha) = \frac{1}{2\widehat{p}}\lambda(1-\widehat{p})\,(e^{-\alpha\eta\widehat{p}} - 1),\\
R_{-+}(\alpha) &= -\frac{1}{2\widehat{p}}\lambda(1+\widehat{p})\,(e^{+\alpha\eta\widehat{p}} - 1), \qquad R_{--}(\alpha) = \frac{1}{2\widehat{p}}\lambda(1-\widehat{p})\,(e^{+\alpha\eta\widehat{p}} - 1),
\end{aligned}
$$

$$
\begin{aligned}
Q_{++}(\alpha) &= \frac{1}{4\widehat{p}^2} e^{-2\alpha\eta\widehat{p}} \Big(2(e^{\alpha\eta\widehat{p}}-1)^2 \lambda^2 (1+\widehat{p}^2) + (e^{2\alpha\eta\widehat{p}}-1)\,\eta\,\widehat{p}\,(v_+(1+\widehat{p}) + v_-(1-\widehat{p})) \Big), \\
Q_{+-}(\alpha) &= \frac{1}{2\widehat{p}^2} e^{-\alpha\eta\widehat{p}} \Big(-(e^{\alpha\eta\widehat{p}}-1)^2 \lambda^2 (1+\widehat{p}^2) - \alpha e^{\alpha\eta\widehat{p}}\,\eta^2\,\widehat{p}^2\,(v_+(1+\widehat{p}) + v_-(1-\widehat{p})) \Big), \\
Q_{--}(\alpha) &= \frac{1}{4\widehat{p}^2} \Big(2(e^{\alpha\eta\widehat{p}}-1)^2 \lambda^2 (1+\widehat{p}^2) + (e^{2\alpha\eta\widehat{p}}-1)\,\eta\,\widehat{p}\,(v_+(1+\widehat{p}) + v_-(1-\widehat{p})) \Big).
\end{aligned}
\tag{37}
$$

The special case of equal prior probabilities is obtained in the limit $\widehat{p} \to 0$:

$$
R_{lm}(\alpha) = l\,m\,\frac{\lambda\eta}{2}\,\alpha \qquad Q_{lm}(\alpha) = l\,m\,\frac{1}{2}\eta^2(\alpha\lambda^2 + v_+ + v_-)\,\alpha. \tag{38}
$$

A.6 The Generalization Error

Using (31) we can directly compute the generalization error as follows:

$$
\varepsilon_g = \sum_{k=\pm 1} p_{-k} \langle \Theta_k \rangle_{-k} = \sum_{k=\pm 1} p_{-k} \Phi\Big(\frac{\widetilde{\beta}_{k-k}}{\widetilde{\alpha}_{k-k}} \Big)
$$

which yields Eqs. (14,15) in the text after inserting
$\widetilde{\alpha}_{sk} = \|C_k^{\frac{1}{2}} \underline{\alpha}_s\| = \sqrt{\underline{\alpha}_s \cdot C_k \underline{\alpha}_s}, \quad \widetilde{\beta}_{sk} = \underline{\alpha}_s \cdot \underline{\mu}_k - \beta_s.$

with $\underline{\alpha}_s = (+2s, -2s, 0, 0)^T$ and $\beta_s = (Q_{+s+s} - Q_{-s-s})$.

References

N. Barkai, H.S. Seung, and H. Sompolinksy. Scaling laws in learning of classification tasks. *Physical Review Letters*, 70(20):3167–3170, 1993.

M. Biehl and N. Caticha. The statistical mechanics of on-line learning and generalization. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1095–1098. MIT Press, Cambridge, MA, 2003.

M. Biehl, A. Freking, and G. Reents. Dynamics of on-line competitive learning. *Europhysics Letters*, 38(1):73–78, 1997.

M. Biehl, A. Ghosh, and B. Hammer. The dynamics of learning vector quantization. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks, ESANN'05*, pages 13–18. d-side, Evere, Belgium, 2005.

T. Bojer, B. Hammer, and C. Koers. Monitoring technical systems with prototype based clustering. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 433–439. d-side, Evere, Belgium, 2003.

L. Bottou. Stochastic gradient learning in neural networks. In *Proc. of Neuro-Nimes 91*. EC2 editions, 1991.

K. Crammer, R. Gilad-Bachrach, A. Navot, and A. Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 462–469. MIT Press, Cambridge, MA, 2003.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, New York, 2000.

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

A. Ghosh, M. Biehl, A. Freking, and 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, available from www.cs.rug.nl/˜biehl*, 2004.

A. Ghosh, M. Biehl, and B. Hammer. Dynamical analysis of LVQ type learning rules. In M. Cottrell, editor, *Workshop on the Self-Organizing-Map WSOM'05*. Univ. de Paris (I), 2005.

B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.

B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Processing Letters*, 21(2):109–120, 2005a.

B. Hammer, M. Strickert, and T. Villmann. Prototype based recognition of splice sites. In U. Seiffert, L. C. Jain, and P. Schweitzer, editors, *Bioinformatics using Computational Intelligence Paradigms*, pages 25–55. Springer, Berlin, 2005b.

B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005c.

T. Kohonen. Learning vector quantization. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks.*, pages 537–540. MIT Press, Cambridge, MA, 1995.

T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1997.

T. Kohonen. Improved versions of learning vector quantization. *In Proc. of the International Joint conference on Neural Networks (San Diego, 1990)*, 1:545–550, 1990.

T. Kohonen, G. Barna, and R. Chrisley. Statistical pattern recognition with neural network: Benchmarking studies. In *Proc. of the IEEE second international conference on Neural Networks (San Diego, 1988)*, volume 1, pages 61–68. IEEE, New York, 1988.

L. I. Kuncheva. Classifier ensembles for changing environments. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems: 5th International Workshop, MCS2004, Cagliari, Italy*, volume 3077 of *Lecture Notes in Computer Science*, pages 1–15. Springer, Berlin, 2004.

C. Marangi, M. Biehl, and S.A. Solla. Supervised learning from clustered input examples. *Europhysics Letters*, 30(2):117, 1995.

R. Meir. Empirical risk minimization versus maximum-likelihood estimation: a case study. *Neural Computation*, 7(1):144–157, 1995.

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

M. Pregenzer, G. Pfurtscheller, and D. Flotzinger. Automated feature selection with distinction sensitive learning vector quantization. *Neurocomputing*, 11:19–20, 1996.

G. Reents and R. Urbanczik. Self-averaging and on-line learning. *Physical Review Letters*, 80(24): 5445–5448, 1998.

P. Riegler, M. Biehl, S.A. Solla, and C. Marangi. On-line learning from clustered input examples. In M. Marinaro and R. Tagliaferri, editors, *Neural Nets WIRN Vietri-95, Proc. of the 7th Italian Workshop on Neural Nets*, pages 87–92. World Scientific, Singapore, 1996.

D. Saad, editor. *Online learning in neural networks*. Cambridge University Press, Cambridge, UK, 1999.

A.S. Sato and K. Yamada. Generalized learning vector quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 423–429, 1995.

A.S. Sato and K. Yamada. An analysis of convergence in generalized LVQ. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *International Conference on Artificial Neural Networks, ICANN'98*, pages 172–176. Springer, Berlin, 1998.

F.-M. Schleif, T. Villmann, and B. Hammer. Local metric adaptation for soft nearest prototype classification to classify proteomic data. In I. Bloch, A. Petrosino, and A.G.B. Tettamanzi, editors, *International Workshop on Fuzzy Logic and Applications*, volume 3849 of *Lecture Notes in Computer Science*, pages 290–296. Springer, Berlin, 2006.

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

S. Seo, M. Bode, and K. Obermayer. Soft nearest prototype classification. *IEEE Transactions on Neural Networks*, 14(2):390–398, 2003.

T. Villmann, E. Merenyi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4):389–403, 2003.

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