Analysis of Robust Soft Learning Vector Quantization

J.J.G. de Vries
J.J.G.de.Vries@student.rug.nl

Institute of Mathematics and Computing Science,
University of Groningen.

Abstract. One of the popular methods for multiclass classification is Learning Vector Quantization (LVQ). There have been developed several variants of LVQ lately, among which Robust Soft Learning Vector Quantization, or RSLVQ for short. An introductory study showed that RSLVQ performs better than other LVQ algorithms, even very close to the optimal linear classifier, within a controlled environment. In order to study its performance in detail, we performed a mathematical analysis of the algorithm, in the form of a system of coupled Ordinary Differential Equations (ODE's), which might also help development of an optimal LVQ algorithm. Following from our analysis, we compare the potential performance of RSLVQ in relation to other LVQ variants and present a guideline for settings of the control parameter, i.e. the softness parameter.

1 Introduction

Learning Vector Quantization (LVQ), originally posed by Kohonen [4, 3] and known by the name LVQ1, is a method of online supervised competitive learning. Many variations on the basic scheme of LVQ1 have been suggested, among which LVQ2.1 and LVQ3 [3, 5], GLVQ [7] and RSLVQ [8, 9] with the aim of obtaining better generalization behavior.

During learning, data samples and their class labels are presented sequentially, or so called 'on-line'. From a set of prototype vectors, defined in the same (potentially high dimensional) space as the data, the closest (set of) prototype(s) is determined and updated such that if the class label coincides with the class label of the data sample, the prototype is attracted to the data, otherwise repelled. The data, carrying labels of different classes, is assumed to be distributed around a specified number of prototypes. Note that there can be more than one prototype per class, enabling a good fit of prototypes to data that contains highly complex class boundaries. Hopefully the prototypes represent the data well after they have settled and training is finished. Classification now can be done by determining the closest of all prototypes and returning the class label corresponding to this winning prototype. Therefore the decision boundaries between the prototypes can also be seen as the Voronoi tessellation of the feature space.

There are a lot of variations of LVQ-algorithms, which mainly differ in which specific prototypes are updated (for example only the closest conflicting prototypes or the closest prototype with corresponding and closest with conflicting label) and how these prototypes are updated. The generic structure of an LVQ algorithm can be expressed in the following way:

$$\begin{aligned} {w_l}^{\mu} &= {w_l}^{\mu-1} + \Delta {w_l}^{\mu}, \\ &= {w_l}^{\mu-1} + \frac{\eta}{N} f(\{{w_i}^{\mu-1}\}, \xi^{\mu}, \sigma^{\mu})(\xi^{\mu} - {w_l}^{\mu-1}) \\ &\quad \text{with } l, i = 1 \ldots c, \mu = 1, 2, \ldots \end{aligned} \tag{1}$$

Where w_l^{μ} is prototype w_l at time step μ, η is the so-called learning rate, N is the dimensionality of the system. The specific form of f is determined by the algorithm used to perform LVQ. The softness of RSLVQ determines the extent to which correctly classified example data (i.e. the closest prototype has a coinciding class label) causes an update of the prototypes. The hard or crisp variant of RSLVQ, in which the softness is taken in the limit to 0, called Learning From Mistakes (LFM) has been analyzed mathematically by Biehl, Ghosh and Hammer [1, 2] in a controlled environment. The study described in this paper is an extension of their study and analyzes truly RSLVQ using the same modelled environment and forms the mathematical background of findings in an introductory study on the performance of RSLVQ [10].

This document is organized as follows: section 2 describes the model in which RSLVQ is analyzed, after which section 3 provides a detailed description of the RSLVQ algorithm. Section 4 describes the analysis globally and section 5 shows the experiments and results of the analysis. Finally section 6 concludes the paper and section 7 gives an overview of future work. The detailed mathematical analysis is attached in appendices A and B.

2 Model

To be able to analyze RSLVQ, we need to restrict the model in which we observe RSLVQ. Biehl et al. [1] defined the following model consisting of high dimensional data originating from a mixture of two overlapping Gaussian clusters, of two classes. We assume that data vectors $\xi \in \mathbb{R}^N$ of class $\sigma \in \{\pm\}$ are drawn independently, with probability $P(\xi)$, according to the following distribution:

$$P(\xi) = \sum_{\sigma=\pm 1} p_{\sigma} P(\xi|\sigma) \tag{2}$$

with

$$P(\xi|\sigma) = \frac{1}{(2\pi v_{\sigma})^{N/2}} \exp\left(\frac{1}{2v_{\sigma}}(\xi - \lambda \boldsymbol{B}_{\sigma})^2\right) \tag{3}$$

The Gaussian clusters are centered around λB_σ with variances v_σ. The prior probabilities p_σ of both classes ($\sigma \in \{+1, -1\}$ or $\{+, -\}$ for short) satisfy $p_+ + p_- = 1$. The vectors B_σ are chosen to be orthonormal, i.e. $B_+^2 = B_-^2 = 1$ and $B_+ \cdot B_- = 0$, so λ specifies the distance (i.e. $\lambda\sqrt{2}$) between the cluster centers. Note that, since λ is chosen such that the clusters overlap, the classification task is clearly not linearly separable. The data points $\{\xi^\mu, \sigma^\mu\}$ are now presented sequentially so that at each time step $\mu = 1, 2, \ldots$ a new uncorrelated vector ξ^μ, along with its label σ^μ, independently drawn according to the density (3), is presented.

There will be fit two prototypes to these clusters, each representing one of the two classes, i.e.:

$$w_s^\mu \in \mathbb{R}^N \text{ with } s \in \{\pm 1\}, \mu = 1, 2, \ldots \tag{4}$$

2.1 Characteristic Quantities

A set of suitable order parameters or characteristic quantities that describe the system has been found by Biehl et al. [1] to be the following:

$$\begin{aligned} R_{S\sigma}^\mu &= \boldsymbol{w}_S^\mu \cdot \boldsymbol{B}_\sigma \\ Q_{ST}^\mu &= \boldsymbol{w}_S^\mu \cdot \boldsymbol{w}_T^\mu \\ &\text{with } \sigma, S, T \in \{\pm 1\}, \mu = 1, 2, \ldots \end{aligned} \tag{5}$$

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. The quantities $R_{+\sigma}^\mu$ and $R_{-\sigma}^\mu$ specify the projections of the prototype vectors into the plane spanned by the vectors B_σ. These characteristic quantities have also been found to express the generalization error in the following way:

$$\begin{aligned} \epsilon_g &= p_+\epsilon_+ + p_-\epsilon_- \\ \epsilon_\sigma &= \phi\Big(\frac{Q_{\sigma\sigma} - Q_{-\sigma-\sigma} - 2\lambda(R_{\sigma\sigma} - R_{-\sigma\sigma})}{2\sqrt{v_\sigma}\sqrt{Q_{++} - 2Q_{+-} + Q_{--}}}\Big) \\ &\text{with } \phi(z) = \int_{-\infty}^{z} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \end{aligned} \tag{6}$$

Note that for all time steps $\mu = 1, 2, \ldots$ these quantities can be determined, resulting in the learning curve or the typical generalization error $\epsilon_g(\alpha)$ after on-line training with $\mu = \alpha N$ random examples.

Details of the calculation can be found, as well as more information on the characteristic quantities, in appendix A.

3 Robust Soft Learning Vector Quantization

As proposed by Seo and Obermayer [9] RSLVQ is a generic algorithm in which different assumptions on the distribution of the data can be made. Note that the assumption about the data distribution to train the prototypes within the LVQ algorithm might differ from the true data distribution used in the model, just as with real world applications in which the true data distribution mostly is not known beforehand. RSLVQ is defined by the following extension of the general update step (1):

$$\begin{aligned} \boldsymbol{w}_{\tilde{l}}^{\mu} &= \boldsymbol{w}_{\tilde{l}}^{\mu-1} + \Delta \boldsymbol{w}_{\tilde{l}}^{\mu} \\ &= \boldsymbol{w}_{\tilde{l}}^{\mu-1} + \tilde{\eta} \begin{cases} \left(P_l(\tilde{l}|\boldsymbol{\xi}^{\mu}) - P(\tilde{l}|\boldsymbol{\xi}^{\mu})\right) \frac{\partial f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})}{\partial \boldsymbol{w}_{\tilde{l}}^{\mu}} & \text{if } l = \sigma^{\mu} \\ -P(\tilde{l}|\boldsymbol{\xi}^{\mu}) \frac{\partial f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})}{\partial \boldsymbol{w}_{\tilde{l}}^{\mu}} & \text{if } l \neq \sigma^{\mu} \end{cases} \end{aligned} \tag{7}$$

where $P_l(\tilde{l}|\boldsymbol{\xi}^{\mu})$ and $P(\tilde{l}|\boldsymbol{\xi}^{\mu})$ are assignment probabilities:

$$\begin{aligned} P_l(\tilde{l}|\boldsymbol{\xi}^{\mu}) &= \frac{p(\tilde{l}) \exp\left(f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})\right)}{\sum_{\tilde{l}=\sigma^{\mu}} p(\tilde{l}) \exp\left(f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})\right)} \\ P(\tilde{l}|\boldsymbol{\xi}^{\mu}) &= \frac{p(\tilde{l}) \exp\left(f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})\right)}{\sum_{\tilde{l}} p(\tilde{l}) \exp\left(f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})\right)} \end{aligned} \tag{8}$$

$P_l(\tilde{l}|\boldsymbol{\xi}^{\mu})$ describes the posterior probability that the data sample $\{\boldsymbol{\xi}^{\mu}, \sigma^{\mu}\}$ is assigned to prototype $\boldsymbol{w}_{\tilde{l}}$ of class l, given that the data sample was generated by the correct class. $P(\tilde{l}|\boldsymbol{\xi}^{\mu})$ describes the posterior probability that the data sample is assigned to prototype $\boldsymbol{w}_{\tilde{l}}$ of all prototypes of all classes. $f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})$ describes the assumed distribution of the data around the prototypes in such a way that $K(\tilde{l})exp\big(f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})\big)$ gives the probability that the data vector $\boldsymbol{\xi}^{\mu}$ is assigned to prototype $\boldsymbol{w}_{\tilde{l}}$.

In our settings we assume a Gaussian distribution, i.e. $K(\tilde{l}) = (2\pi v_{\tilde{l}})^{N/2}$ and $f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu}) = -(\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu})^2/2v_{\tilde{l}}$, implying $\frac{\partial f(\boldsymbol{\xi}, \boldsymbol{w}_{\tilde{l}}^{\mu})}{\partial \boldsymbol{w}_{\tilde{l}}^{\mu}} = \frac{\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu}}{v_{\tilde{l}}}$, and the prototypes $\boldsymbol{w}_S$ all have the same width and strength, i.e. the variances and priors are equal: $\forall S :: v_S = v_{soft}, p(S) = \frac{1}{\#\boldsymbol{w}_S}$.

With these assumptions the update rule (7) becomes:

$$\boldsymbol{w}_{\tilde{l}}^{\mu} = \boldsymbol{w}_{\tilde{l}}^{\mu-1} + \frac{\tilde{\eta}}{v_{soft}} \begin{cases} \left(P_l(\tilde{l}|\boldsymbol{\xi}^{\mu}) - P(\tilde{l}|\boldsymbol{\xi}^{\mu})\right)(\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu}) & \text{if } l = \sigma^{\mu} \\ -P(\tilde{l}|\boldsymbol{\xi}^{\mu})(\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu}) & \text{if } l \neq \sigma^{\mu} \end{cases} \tag{9}$$

where

$$P_l(\tilde{l}|\boldsymbol{\xi}^{\mu}) = \frac{\exp\left(-(\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu})^2/2v_{soft}\right)}{\sum_{\tilde{l}=\sigma^{\mu}} \exp\left(-(\boldsymbol{\xi} - \boldsymbol{w}_{\tilde{l}}^{\mu})^2/2v_{soft}\right)}$$

$$P(\tilde{l}|\xi^{\mu}) = \frac{\exp\big(-(\xi - w_{\tilde{l}}^{\mu})^2/2v_{soft}\big)}{\sum_{\tilde{l}}\exp\big(-(\xi - w_{\tilde{l}}^{\mu})^2/2v_{soft}\big)} \tag{10}$$

Since in our model only two prototypes, each representing a different class, are used, equation (10) can be written as:

$$\begin{aligned} P_l(l|\xi^{\mu}) &= \frac{\exp\big(-(\xi - w_l^{\mu})^2/2v_{soft}\big)}{\exp\big(-(\xi - w_l^{\mu})^2/2v_{soft}\big)} = 1 \\ P(l|\xi^{\mu}) &= \frac{\exp\big(-(\xi - w_l^{\mu})^2/2v_{soft}\big)}{\exp\big(-(\xi - w_+^{\mu})^2/2v_{soft}\big) + \exp\big(-(\xi - w_-^{\mu})^2/2v_{soft}\big)} \\ &= \frac{1}{1 + \exp\Big(\big((\xi - w_l^{\mu})^2 - (\xi - w_{-l}^{\mu})^2\big)/2v_{soft}\Big)} \end{aligned} \tag{11}$$

Putting this back into equation (9) yields:

$$\begin{aligned} w_l{}^{\mu} &= w_l{}^{\mu-1} + \frac{\tilde{\eta}}{v_{soft}}\Big(\delta_{l\sigma^{\mu}} - \frac{1}{1 + \exp(\frac{d_l^{\mu} - d_{-l}^{\mu}}{2v_{soft}})}\Big)(\xi^{\mu} - w_l{}^{\mu-1}) \\ &\text{with } d_l^{\mu} = (\xi - w_l^{\mu})^2 \end{aligned} \tag{12}$$

It is however more convenient to rescale the learning rate with the dimensionality of the data (N), so we rewrite:

$$w_l{}^{\mu} = w_l{}^{\mu-1} + \frac{\eta}{Nv_{soft}}(\delta_{l\sigma^{\mu}} - \Omega_l)(\xi^{\mu} - w_l{}^{\mu-1}) \tag{13}$$

Where $\delta_{l\sigma^{\mu}}$ is the Kronecker delta and

$$\Omega_l = \frac{1}{1 + \exp(\frac{d_l^{\mu} - d_{-l}^{\mu}}{2v_{soft}})} \tag{14}$$

However, in the form of equation (12), the update function cannot be integrated analytically, as we would like in the further analysis. As an alternative route we approximate the update (12) by an LVQ variant which facilitates further analytic treatment. We use the observation that $\frac{1}{1+\exp(x)}$ is very similar to $\Phi(\frac{-x}{c})$, where $c \in \mathbb{R}$ is a constant which controls the slope of the Φ-function and we rewrite:

$$w_l{}^{\mu} = w_l{}^{\mu-1} + \frac{\eta}{Nv_{soft}}(\delta_{l\sigma^{\mu}} - \Phi_l)(\xi^{\mu} - w_l{}^{\mu-1}) \tag{15}$$

Where

$$\Phi_l = \Phi\Big(\frac{d_{-l}^{\mu} - d_l^{\mu}}{2cv_{soft}}\Big) \tag{16}$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \tag{17}$$

To obtain the value of c let us set equal the slope at $x = 0$, therefore observe the derivative of both activation functions:

$$\begin{aligned} \frac{d}{dx}\frac{1}{1+e^x} &= -\frac{e^x}{(1+e^x)^2} \\ \frac{d}{dx}\Phi(\frac{-x}{c}) &= -\frac{e^{-\frac{x^2}{2c^2}}}{c\sqrt{2\pi}} \end{aligned} \tag{18}$$

Now set these derivatives equal and plug in $x = 0$:

$$\begin{aligned} -\frac{e^0}{(1+e^0)^2} &= -\frac{e^{-\frac{0}{2c^2}}}{c\sqrt{2\pi}} \\ -\frac{1}{4} &= -\frac{1}{c\sqrt{2\pi}} \\ c &= \frac{4}{\sqrt{2\pi}} \end{aligned} \tag{19}$$

4 Analysis

The mathematical analysis of RSLVQ consists of the following steps:

- Describe development of the characteristic quantities in terms of recurrence relations
- Turning of recurrence relations into differential equations
- Performing averages on the differential equations

These three steps are described in appendix B and give us mathematical descriptions of the development of the system with learning time. There are multiple variants of the differential equations. First of all there are two variants of RSLVQ: original (equation (13)) and with Φ-approximation (equation(15)), both in the limit $\eta \to 0$. In this limit it is possible to analytically determine the ODE's for Φ-approximated RSLVQ while the ODE's for original RSLVQ contain numerical integrations.

The most interesting results are the stationary states of the system of coupled ODE's. These states can be obtained using large learning times or, more reliably, by searching for zeros of the right hand sides of the ODE's. To search for zeros of a seven-dimensional non-linear system is, however, difficult, therefore we search for zeros in the sum of squared right hand side terms, with the restriction that the solution is physical, i.e. the covariance matrix C_k, see equation (34), should be positive semi-definite, which is the case when all its eigenvalues are non-negative.

For determining optimal settings of the control parameter v_{soft} we interpret the generalization error in the stationary states as a function of v_{soft}. This function can then be used to find the optimal setting by searching for the minimum.

5 Experiments and Results

Several experiments have been conducted, some of them containing a comparison of simulations and differential equations. Note that the differential equations are determined in the limit $N \to \infty$. The simulations were performed with $N = 100$, which is obviously sufficient to match the theory for $N \to \infty$, see [1] for a discussion of finite N corrections.

5.1 Simulations versus ODE

First we will show that the ODE's indeed describe the system by comparison of the development of the characteristic quantities of simulated training with those of the ODE' s.

Original RSLVQ As one can see from figure 1, the ODE's match the simulations exactly, for original RSLVQ. Note that the learning rate of $\eta = 0.05$ already is small enough for the simulations to match the ODE's which are valid in the limit $\eta \to 0$. The softness v_{soft} is chosen such that it is not too small and not too large because it has a similar effect as the learning rate: too large makes the system taking too big steps resulting in poor convergence and too small causes the system to converge very slowly and besides that it shows the limiting behavior of LFM, i.e. poor performance and instability issues.

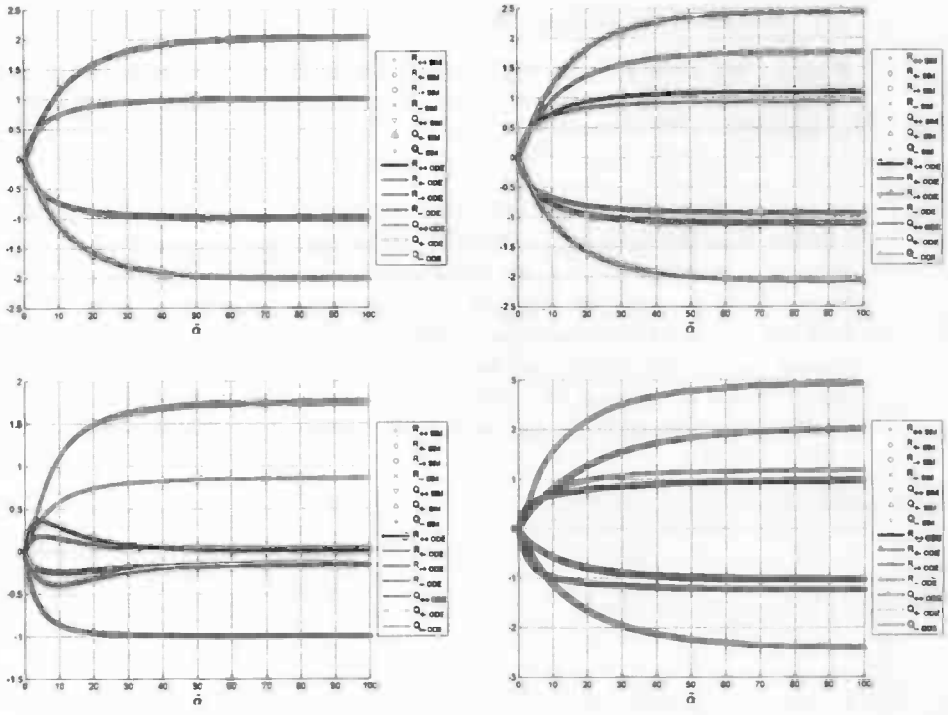

Fig. 1. Development of the characteristic quantities with the learning time $\tilde{\alpha}$ of the original RSLVQ for both simulations (thick lines) and ODE's (thinner lines that lay on top of the thick lines) for different systems. Top left: $p_+ = p_- = 0.5$, $v_+ = v_- = 0.5$; top right: $p_+ = p_- = 0.5$, $v_+ = 0.25$, $v_- = 0.81$; bottom left: $p_+ = 0.7$, $v_+ = v_- = 1$; bottom right: $p_+ = 0.7$, $v_+ = 0.25$, $v_- = 0.81$. The softness v_{soft} has been set to 1 and the learning rate used in the simulations is $\eta = 0.05$.

RSLVQ with Φ-approximation The same experiment has also been conducted with RSLVQ with Φ-approximation for both simulations and ODE's.

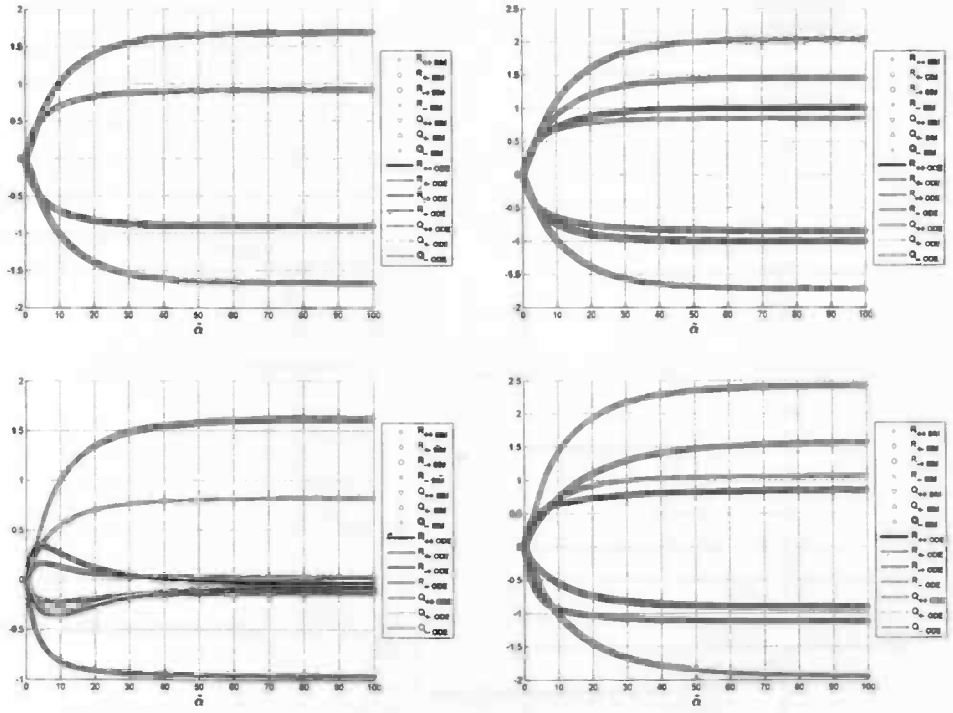

Fig. 2. Development of the characteristic quantities with the learning time $\tilde{\alpha}$ of RSLVQ with Φ-approximation for both simulations (thick lines) and ODE's (thinner lines that lay on top of the thick lines) for different systems. Top (bottom): (un)equal prior and left (right): (un)equal data variances; See figure 1 for detailed information on the settings.

As one can see from figure 2, the ODE's match the simulations exactly, for Φ-approximated RSLVQ as well.

5.2 Comparison

Let us now compare the original with the Φ-approximated version of RSLVQ:

As one can see from figure 3, the Φ-approximation does influence the development of the characteristic quantities, i.e. the ODE's (and therefore the simulations) differ from original RSLVQ and Φ-approximated RSLVQ, however the tendency of each of the quantities is the same and deviations are not too

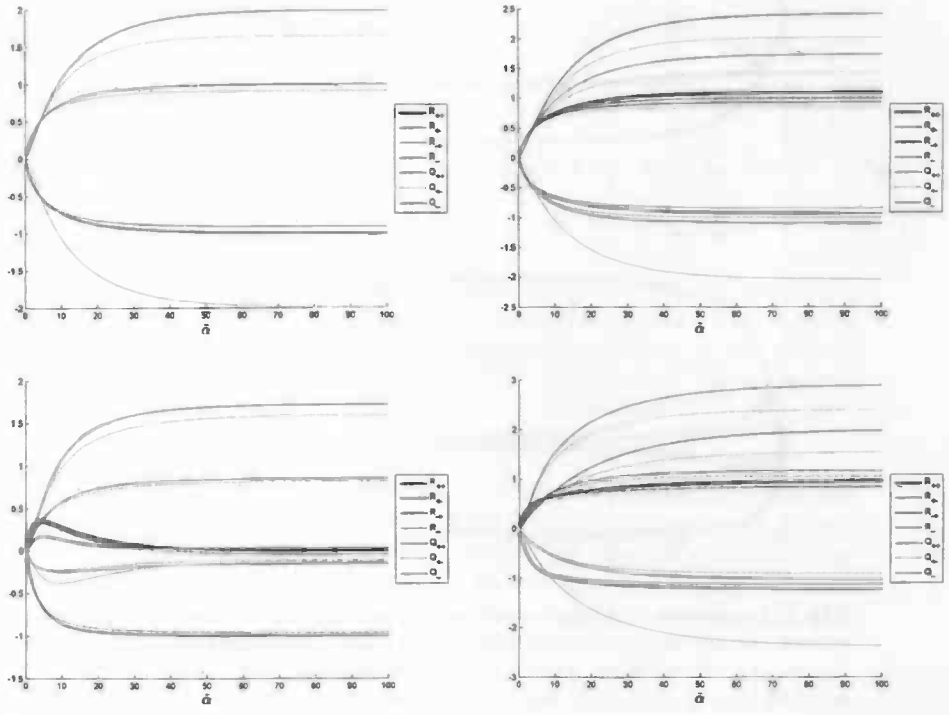

Fig. 3. Comparison of the characteristic quantities with the learning time $\tilde{\alpha}$ of RSLVQ with (dotted) and without (solid) Φ-approximation for different systems. Top (bottom): (un)equal prior and left (right): (un)equal data variances; See figure 1 for detailed information on the settings.

large. Moreover the generalization ability is not affected as can be seen in figure 4, which shows a perfect match of the developement of the generalization error during learning for original and Φ-approximated RSLVQ. Since the generalization ability is the main target of our study, we conclude that the ODE's for Φ-approximated RSLVQ describe the training process of the original RSLVQ algorithm well and can be used to study the algorithms performance. Note however that these ODE's are only valid for small η, i.e. in the limit $\eta \to 0$.

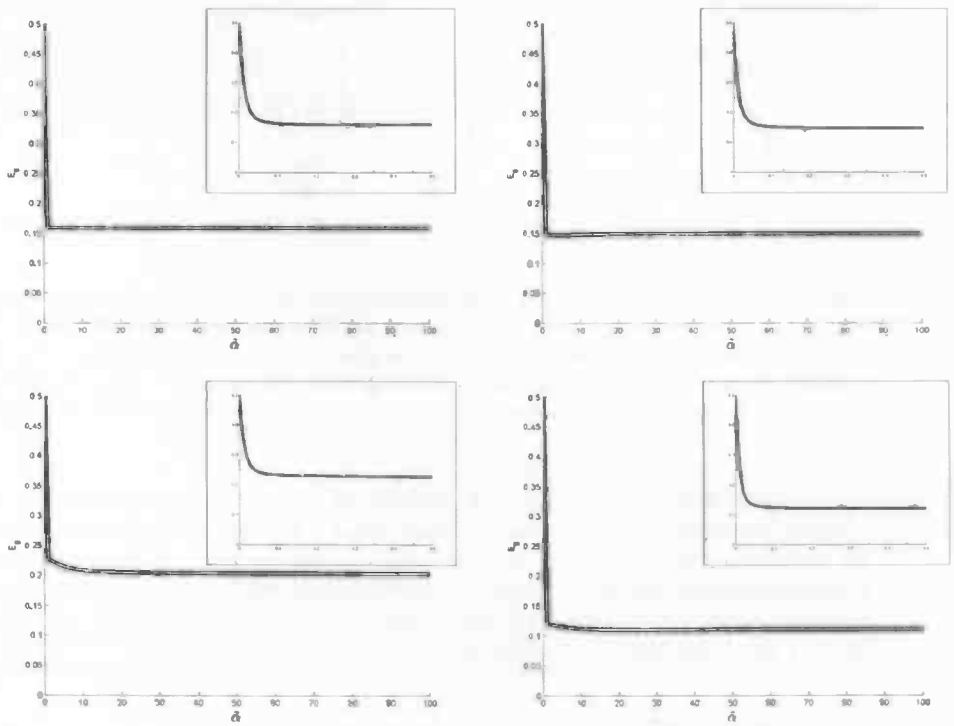

Fig. 4. Comparison of the generalization error (E_g) with the learning time $\tilde{\alpha}$ of RSLVQ with (small red) and without (thick black) Φ-approximation for different systems. Top (bottom): (un)equal prior and left (right): (un)equal data variances; See figure 1 for detailed information on the settings. The insets show a close up of the first part of the graphs, i.e. $\tilde{\alpha} \leq 0.5$.

5.3 Asymptotic Performance

The performance of the LVQ algorithms is measured in terms of the generalization error in the stationary states. These states correspond to zeros in the

derivatives of the ODE's. Because it is difficult to search for the zeros of 7 coupled ODE's, we search for the zeros of the sum of quadratic right hand side terms of the ODE's. The optimal generalization error can be determined for all settings of priors per setting of the data variances.

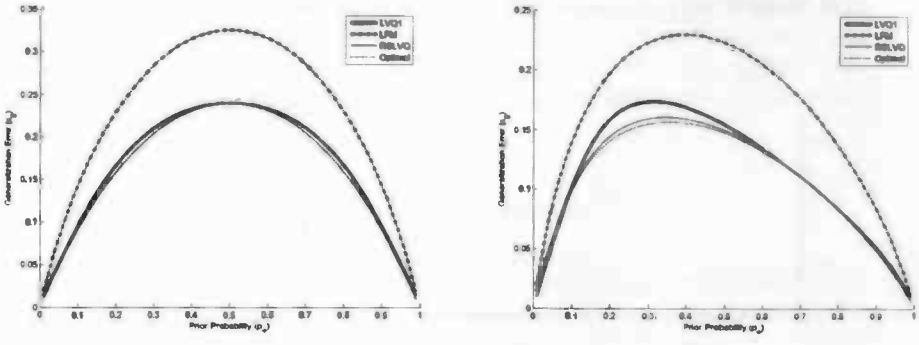

Fig. 5. Asymptotic performance, i.e. the generalization error in the stationary states, of RSLVQ in comparison with other LVQ algorithms. The dashed (lowest) line marks the best linear classifier, the continuous black line marks LVQ1, LFM is marked with the chained line and the red line represents RSLVQ, with the optimal or close to optimal choices of the softness parameter for left: $v_+ = v_- = 1$ and right: $v_+ = 0.25, v_- = 0.81$.

As figure 5 shows, RSLVQ with optimal or close to optimal choices of the softness parameter performs well beyond other LVQ algorithms, even optimal for equal data variances and very close to optimal for unequal data variances. Perhaps this last bit of performance can be gained by using multiple softness parameters, i.e. one per class or per prototype, to enable the system to fit fully to data with unequal data variances.

5.4 Optimal softness

One can write the limiting performance (generalization error) as a function of the softness parameter (v_{soft}) and use this function to numerically find the minimum and therefore the optimal setting of the softness parameter. This has been done for several settings, of which table 1 gives an overview.

As one can see from table 1, the optima found by numerical search are close to the maximum of data variances ($\max(v_+, v_-)$). However, to interpret these numbers correctly, let us look at how the function behaves in the surrounding area.

As one can see from figure 6, there is a large flat region in the generalization error for small values of v_{soft} for both equal and unequal priors. There is presumably a single optimal value, but the true minimum is apparently very flat, yielding very robust algorithm with respect to misestimation of the softness.

	$v_+ = 0,25$ $v_- = 0,25$	$v_+ = 0,5$ $v_- = 0,5$	$v_+ = 1$ $v_- = 1$	$v_+ = 0,25$ $v_- = 0,81$
$p_+ = 0,5$	0,3125	0,51094	1,0625	0,78722
$p_+ = 0,7$	0,25156	0,51328	0,98145	0,83025

Table 1. Optimal settings of v_{soft}, found by numerical search for minima of the generalization error.

It is however possible that for some settings of the softness parameter the (close to) optimal generalization error is reached earlier during training than for others. Either a too small or too large value of v_{soft} results in slow or poor convergence. Furthermore a too small softness (in the limit $v_{soft} \to 0$) will reach the limiting behavior of LFM, which has stability issues.

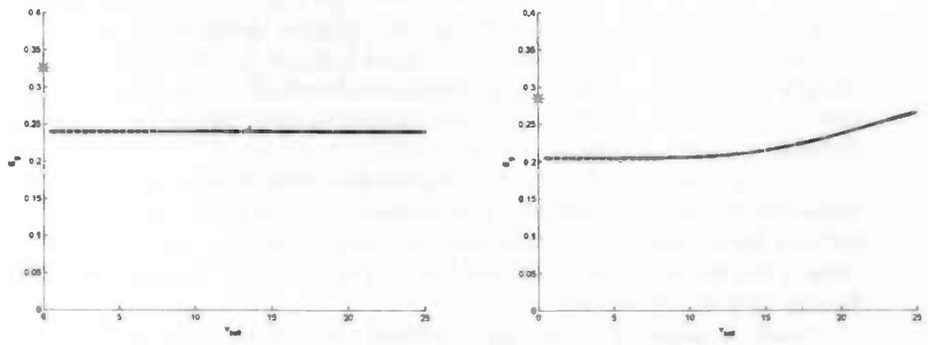

Fig. 6. Dependence of the generalization error for large $\tilde{\alpha}$ on v_{soft} with left: $p_+ = 0.5$ and right: $p_+ = 0.7$, $v_+ = v_- = 1$, showing robustness with respect to settings of the softness parameter. The red star marks the generalization error for LFM, however with finite η because for LFM the limit $\eta \to 0$ causes instability.

Note that different settings of the data variances showed similar graphs, however for smaller data variances the extension of the close to optimal range of softness decreases. Taking this into consideration, the expectation is that a setting of $1 \leq v_{soft} \leq 2$ will give good performance within reasonable training time for most applications.

6 Discussion

We showed that we can mathematically describe the learning behavior of RSLVQ by using a system of 7 coupled ordinary differential equations. In the limit $\eta \to 0$ we can even calculate them analytically for RSLVQ with Φ-approximation. Without Φ-approximation we encounter numerical integrations. The ODE's describe

the system well since they fit the to the simulations, for both original and Φ-approximated RSLVQ. The Φ-approximation does influence the behavior of the system, however the quantity of most interest, the generalization error, is not affected by the approximation.

The performance of RSLVQ is well beyond other LVQ algorithms, confirming the findings of the introductory study that was based on simulations only, and even optimal for equal data variances and close to optimal for unequal data variances. Finally we saw that there is presumably a single optimal setting of the softness parameter, but there is also a flat region of close to optimal softness, enabling an easy choice for setting this control parameter for practical applications.

7 Future work

We were able to describe the learning behavior of RSLVQ mathematically in the limit $\eta \to 0$. ODE's for finite η contain large numerical integrations, causing large time needed for solving the ODE's. It would however be interesting to compare the results for finite η with the results we found. Perhaps some simplifying assumptions enable to get rid of some numerical integrations in the ODE's for finite η.

RSLVQ turned out to perform slightly less then optimal for unequal data variances. Perhaps this last bit of performance can be gained by using multiple softness parameters, for example one per class or one per prototype (this coincides in our model). This could well lead to the optimal LVQ algorithm, at least for the data model we used.

Finally it might be interesting to extend the analysis to more than two prototypes since RSLVQ cannot show its full potential in our model yet, because of the choice for using only two prototypes. By using more than two prototypes per class, the algorithm is able to fit to nonlinear decision boundaries. This however would imply an elaborate extension of the calculations because our analysis is based on using two prototypes.

Acknowledgements

First and foremost I would like to thank my supervisor, and leader of the *LVQ research group*, Michael Biehl for his guidance throughout my master thesis project. He was always available to answer my questions. I am also very grateful to Anarta Ghosh, who has helped me a lot with the calculations.

I would like to thank Aree Witoelar, member of the *LVQ research group*, for his help on debugging the calculations and code for solving the differential equations. The biweekly meetings of the *LVQ research group* have also provided me more insight in various problems and worked as inspiration for my project.

Finally I would like to express my gratitude for the support from my family during the project.

References

1. M. Biehl, A. Ghosh and B. Hammer, *Dynamics and generalization ability of LVQ algorithms,* in Journal of Machine Learning Research 8 (Feb):323-360, 2007.
2. A. Ghosh, M. Biehl and B. Hammer, *Dynamical analysis of LVQ type learning rules,* Workshop on the Self-Organizing-Map, WSOM'05, 2005.
3. T. Kohonen, *Improved versions of learning vector quantization,* Procedings of the International Joint conference on Neural Networks (Sand Diego, 1990), 1:545-550, 1990.
4. T. Kohonen, *Learning vector quantization,* M. Arbib, editor, The handbook of brain theory and neural networks, 537-540, MIT Press, Cambridge, MA, 1995.
5. T. Kohonen, *Self-Organizing Maps,* Springer, Berlin, 1997.
6. G. Reents and R. Urbanczik, *Self-Averaging and On-Line Learning,* Physical review letters. Vol. 80. No. 24, pp. 5445-5448, 1998.
7. 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, 423-429, 1995.
8. S. Seo, M. Bode and K. Obermayer, *Soft nearest prototype classification,* IEEE Transactions on Neural Networks, 13(2):390-398, 2003.
9. S. Seo and K. Obermayer, *Soft learning vector quantization,* Neural computation, 15: 1589-1603, 2003.
10. J.J.G. de Vries, *The behaviour of RSLVQ in a controlled environment,* Rijksuniversiteit Groningen, internal document, 2006.

A Statistics of the model

A.1 Notations

Let us first introduce the following notations:

For any $x \in \mathbb{R}^N$, $x^2 \equiv x \cdot x$; $\cdot$ denotes scalar product. $< . >$ denotes the average (expectation) over $p(\xi)$ and can be expressed in the following form:

$$< . >= \sum_{\sigma=\pm 1} p_\sigma < . >_\sigma \tag{20}$$

Where $< . >_\sigma$ is the conditional average for class σ.

A.2 Statistics of the data

From equation (3) follows that the components ξ_j are statistically independent, Gaussian distributed, quantities with variance v_σ and mean $< \xi_j >_\sigma = \lambda(B_\sigma)_j$. Furthermore, note that for a statistical quantity $X \sim N(\mu, \sigma)$ it holds that $\sigma^2 =< X^2 > - < X >^2$, so $< X^2 >=< X >^2 +\sigma^2$. It follows that:

$$\begin{aligned}
< \xi^2 >_\sigma &= \sum_{j=1}^{N} < \xi_j^2 >_\sigma \\
&= \sum_{j=1}^{N} (v_\sigma + < \xi_j >^2) \\
&= \sum_{j=1}^{N} (v_\sigma + (\lambda(B_\sigma)_j)^2) \\
&= v_\sigma N + \lambda^2 \sum_{j=1}^{N} (B_\sigma)_j^2 \\
&= v_\sigma N + \lambda^2
\end{aligned} \tag{21}$$

Note that in the last step it is used that $\sum_{j=1}^{N} (B_\sigma)_j^2 = 1$.

Thus we obtain:

$$\begin{aligned}
< \xi \cdot \xi > &= \sum_{\sigma=\pm 1} p_\sigma < \xi^2 >_\sigma \\
&= p_{+1}(N v_{+1} + \lambda^2) + p_{-1}(N v_{-1} + \lambda^2) \\
&= N(p_{+1} v_{+1} + p_{-1} v_{-1}) + \lambda^2 \\
&\approx N(p_{+1} v_{+1} + p_{-1} v_{-1}) \\
&\quad \left[\because N >> \lambda \right]
\end{aligned} \tag{22}$$

A.3 Order Parameters and Projections

Define the order parameters (R_{lm}, Q_{lm}) and the projections (b_l, h_l) as follows:

$$\begin{aligned} R_{lm} &= \boldsymbol{w}_l \cdot \boldsymbol{B}_m \\ Q_{lm} &= \boldsymbol{w}_l \cdot \boldsymbol{w}_m \\ b_l &= \boldsymbol{\xi} \cdot \boldsymbol{B}_l \\ h_l &= \boldsymbol{\xi} \cdot \boldsymbol{w}_l \end{aligned} \tag{23}$$

Define,

$$\boldsymbol{x} = (h_1, h_{-1}, b_1, b_{-1}) \tag{24}$$

A.4 Statistics of the Projections

Given that each training vector is independent of all previous ones, the statistical properties of the projections are well defined for large N. The central limit theorem yields that their joint density, $p(h_{+1}, h_{-1}, b_{+1}, b_{-1}) = p(\boldsymbol{x})$, is normally distributed and fully specified by the corresponding conditional averages and covariances.

First Order Statistics of $\boldsymbol{h}$:

$$\begin{aligned} < h_l >_k &= < w_l \cdot \boldsymbol{\xi} >_k \\ &= \boldsymbol{w}_l \cdot < \boldsymbol{\xi} >_k \\ &= \boldsymbol{w}_l \cdot \lambda \boldsymbol{B}_k \\ &= \lambda R_{lk} \end{aligned} \tag{25}$$

First Order Statistics of $\boldsymbol{b}$:

$$\begin{aligned} < b_l >_k &= < \boldsymbol{B}_l \cdot \boldsymbol{\xi} >_k \\ &= \boldsymbol{B}_l \cdot < \boldsymbol{\xi} >_k \\ &= \boldsymbol{B}_l \cdot \lambda \boldsymbol{B}_k \\ &= \begin{cases} \lambda \text{ if } l = k, \text{ note } B_l^2 = 1 \\ 0 \text{ if } l \neq k, \text{ note } B_l \cdot B_k = 0 \end{cases} \\ &= \lambda \delta_{lk} \end{aligned} \tag{26}$$

Where δ_{lk} is the Kronecker delta. Hence the conditional means of $\boldsymbol{x}$ for two classes can be expressed in the following way:

$$\boldsymbol{\mu}_{+1} = \begin{pmatrix} \lambda R_{+1,+1} \\ \lambda R_{-1,+1} \\ 1 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\mu}_{-1} = \begin{pmatrix} \lambda R_{+1,-1} \\ \lambda R_{-1,-1} \\ 0 \\ 1 \end{pmatrix} \tag{27}$$

Second Order Statistics of h: $< h_l h_m >_k - < h_l >_k < h_m >_k$
To compute the conditional variance let us first look at the average,

$$
\begin{aligned}
< h_l h_m >_k &= < (\boldsymbol{w}_l \cdot \boldsymbol{\xi})(\boldsymbol{w}_m \cdot \boldsymbol{\xi}) >_k \\
&= \left\langle \left(\sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{\xi})_i \right) \left(\sum_{j=1}^{N} (\boldsymbol{w}_m)_j (\boldsymbol{\xi})_j \right) \right\rangle_k \\
&= \sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_i < (\boldsymbol{\xi})_i^2 >_k + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_j < (\boldsymbol{\xi})_i (\boldsymbol{\xi})_j >_k \\
&= \sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_i \left[v_k + \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_i \right] \\
&\quad + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_j \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&\quad \left[\because \begin{array}{l} \text{components of } \boldsymbol{\xi} \text{ have variance } v_k, \text{ also see equation (21),} \\ \text{and are independent} \end{array} \right] \\
&= v_k \sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_i + \lambda^2 \sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_i (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_i \\
&\quad + \lambda^2 \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_j (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{w}_l \cdot \boldsymbol{w}_m + \lambda^2 \sum_{i=1}^{N} \sum_{j=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{w}_m)_j (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{w}_l \cdot \boldsymbol{w}_m + \lambda^2 (\boldsymbol{w}_l \cdot \boldsymbol{B}_k)(\boldsymbol{w}_m \cdot \boldsymbol{B}_k) \\
&= v_k Q_{lm} + \lambda^2 R_{lk} R_{mk} \qquad (28)
\end{aligned}
$$

Hence we have,

$$
\begin{aligned}
< h_l, h_m >_k - < h_l >_k < h_m >_k &= v_k Q_{lm} + \lambda^2 R_{lk} R_{mk} - \lambda^2 R_{lk} R_{mk} \\
&= v_k Q_{lm} \qquad (29)
\end{aligned}
$$

Second Order Statistics of b: Similar to equation (28) we get the second order statistics for b as follows:

$$
\begin{aligned}
< b_l b_m >_k &= < (\boldsymbol{B}_l \cdot \boldsymbol{\xi})(\boldsymbol{B}_m \cdot \boldsymbol{\xi}) >_k \\
&= \left\langle \left(\sum_{i=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{\xi})_i \right) \left(\sum_{j=1}^{N} (\boldsymbol{B}_m)_j (\boldsymbol{\xi})_j \right) \right\rangle_k \\
&= \sum_{i=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_i < (\boldsymbol{\xi})_i^2 >_k + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_j < (\boldsymbol{\xi})_i (\boldsymbol{\xi})_j >_k
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_i \left[v_k + \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_i \right] \\
&\quad + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_j \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&\quad \left[\because \begin{array}{l} \text{components of } \xi \text{ have variance } v_k, \text{ also see equation (21),} \\ \text{and are independent} \end{array} \right] \\
&= v_k \sum_{i=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_i + \lambda^2 \sum_{i=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_i (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_i \\
&\quad + \lambda^2 \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_j (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{B}_l \cdot \boldsymbol{B}_m + \lambda^2 \sum_{i=1}^{N} \sum_{j=1}^{N} (\boldsymbol{B}_l)_i (\boldsymbol{B}_m)_j (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{B}_l \cdot \boldsymbol{B}_m + \lambda^2 (\boldsymbol{B}_l \cdot \boldsymbol{B}_k)(\boldsymbol{B}_m \cdot \boldsymbol{B}_k) \\
&= \delta_{lm} v_k + \lambda^2 \delta_{lk} \delta_{mk} \\
&\quad \left[\because B_l \cdot B_k = \begin{cases} 1 \text{ if } l = k \\ 0 \text{ if } l \neq k \end{cases} \right] \\
&= \delta_{lm} (v_k + \lambda^2 \delta_{lk}) \qquad (30)
\end{aligned}
$$

Hence we have,

$$
\begin{aligned}
< b_l b_m >_k - < b_l >_k < b_m >_k &= \delta_{lm} (v_k + \lambda^2 \delta_{lk}) - \lambda^2 \delta_{lk} \delta_{mk} \\
&= \delta_{lm} v_k \qquad (31)
\end{aligned}
$$

Covariance of h and b: To compute the conditional variance, $< h_l b_m >_k - < h_l >_k < b_m >_k$, let us first look at the average,

$$
\begin{aligned}
< h_l b_m >_k &= < (w_l \cdot \xi)(\boldsymbol{B}_m \cdot \xi) >_k \\
&= \left\langle \left(\sum_{i=1}^{N} (w_l)_i (\xi)_i \right) \left(\sum_{j=1}^{N} (\boldsymbol{B}_m)_j (\xi)_j \right) \right\rangle_k \\
&= \left\langle \sum_{i=1}^{N} (w_l)_i (\boldsymbol{B}_m)_i (\xi)_i (\xi)_i + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (w_l)_i (B_m)_j (\xi)_i (\xi)_j \right\rangle_k \\
&= \sum_{i=1}^{N} (w_l)_i (\boldsymbol{B}_m)_i < (\xi)_i (\xi)_i >_k \\
&\quad + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (w_l)_i (\boldsymbol{B}_m)_j < (\xi)_i (\xi)_j >_k
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} (\boldsymbol{w}_l)_i (\boldsymbol{B}_m)_i \left[v_k + \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_i \right] \\
&+ \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (w_l)_i (\boldsymbol{B}_m)_j \lambda^2 (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{w}_l \cdot \boldsymbol{B}_m + \lambda^2 \sum_{i=1}^{N} \sum_{j=1}^{N} (w_l)_i (\boldsymbol{B}_m)_j (\boldsymbol{B}_k)_i (\boldsymbol{B}_k)_j \\
&= v_k \boldsymbol{w}_l \cdot \boldsymbol{B}_m + \lambda^2 (\boldsymbol{w}_l \cdot \boldsymbol{B}_k)(\boldsymbol{B}_m \cdot \boldsymbol{B}_k) \\
&\quad \left[\because B_m \cdot B_k = \begin{cases} 1 \text{ if } m = k \\ 0 \text{ if } m \neq k \end{cases} \right] \\
&= v_k R_{lm} + \lambda^2 R_{lk} \delta_{mk} \qquad (32)
\end{aligned}
$$

Hence we have,

$$
\begin{aligned}
< h_l, b_m >_k - < h_l >_k < b_m >_k &= v_k R_{lm} + \lambda^2 R_{lk} \delta_{mk} - \lambda R_{lk} \lambda \delta_{mk} \\
&= v_k R_{lm} \qquad (33)
\end{aligned}
$$

The conditional density of x for class k is $N(\mu_k, \boldsymbol{C}_k)$, where, μ_k is the conditional mean vector for class k and $\boldsymbol{C}_k$ is the conditional covariance matrix for class k. In our model the conditional density of x is a 4-dimensional Gaussian. Following from equations (29, 31, 33), the covariance matrix by $\boldsymbol{C}_k$ and can be expressed as follows:

$$
\boldsymbol{C}_k = v_k \begin{pmatrix} Q_{+1,+1} & Q_{+1,-1} & R_{+1,+1} & R_{+1,-1} \\ Q_{+1,-1} & Q_{-1,-1} & R_{-1,+1} & R_{-1,-1} \\ R_{+1,+1} & R_{-1,+1} & 1 & 0 \\ R_{+1,-1} & R_{-1,-1} & 0 & 1 \end{pmatrix} \qquad (34)
$$

B Detailed derivation of system of coupled ODE's

Because each training sample that is presented is independent of its predecessors we can conceptually think of training as a stochastic process, to be precise a Markov process (i.e. the future states do depend on the current state but are independent of previous states). If the underlying distribution of the training data is simple enough, the whole dynamics of the system can be analyzed using a few characteristic quantities $\{R_{lm}, Q_{lm}\}$. These order parameters are self-averaging [6] in the thermodynamic limit ($N \to \infty$), allowing to analyze the stochastic evolution of the system in terms of deterministic evolution of the characteristic quantities. The evolution of characteristic quantities is described by a system of coupled differential equations.

B.1 Recurrence relation for the characteristic quantities

Recurrence relation for R: From equation (15) follows:

$$\begin{aligned}
R_{lm}^{\mu} &= \boldsymbol{w}_l^{\mu} \cdot \boldsymbol{B}_m \\
&= \left(\boldsymbol{w}_l^{\mu-1} + \Delta \boldsymbol{w}_l{}^{\mu}\right) \cdot \boldsymbol{B}_m \\
&= R_{lm}^{\mu-1} + \Delta \boldsymbol{w}_l{}^{\mu} \cdot \boldsymbol{B}_m \\
&= R_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} (\delta_{l\sigma^{\mu}} - \Phi_l)(\boldsymbol{\xi}^{\mu} - \boldsymbol{w}_l{}^{\mu-1}) \cdot \boldsymbol{B}_m \\
&= R_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} \Big(\delta_{l\sigma^{\mu}} \boldsymbol{\xi}^{\mu} - \delta_{l\sigma^{\mu}} \boldsymbol{w}_l{}^{\mu-1} - \Phi_l \boldsymbol{\xi}^{\mu} \\
&\quad + \Phi_l \boldsymbol{w}_l{}^{\mu-1}\Big) \cdot \boldsymbol{B}_m \\
&= R_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} \Big(\delta_{l\sigma^{\mu}} b_m^{\mu} - \delta_{l\sigma^{\mu}} R_{lm}^{\mu-1} - \Phi_l b_m^{\mu} + \Phi_l R_{lm}^{\mu-1}\Big) \qquad (35)
\end{aligned}$$

Recurrence relation for Q:

$$\begin{aligned}
Q_{lm}^{\mu} &= \boldsymbol{w}_l^{\mu} \cdot \boldsymbol{w}_m^{\mu} \\
&= \left(\boldsymbol{w}_l{}^{\mu-1} + \Delta \boldsymbol{w}_l{}^{\mu}\right) \cdot \left(\boldsymbol{w}_m{}^{\mu-1} + \Delta \boldsymbol{w}_m{}^{\mu}\right) \\
&= \boldsymbol{w}_l{}^{\mu-1} \cdot \boldsymbol{w}_m{}^{\mu-1} + \boldsymbol{w}_l{}^{\mu-1} \cdot \Delta \boldsymbol{w}_m{}^{\mu} + \Delta \boldsymbol{w}_l{}^{\mu} \cdot \boldsymbol{w}_m{}^{\mu-1} + \Delta \boldsymbol{w}_l{}^{\mu} \cdot \Delta \boldsymbol{w}_m{}^{\mu} \\
&= Q_{lm}^{\mu-1} + \boldsymbol{w}_l{}^{\mu-1} \cdot \left(\frac{\eta}{N v_{soft}} (\delta_{m\sigma^{\mu}} - \Phi_m)(\boldsymbol{\xi}^{\mu} - \boldsymbol{w}_m{}^{\mu-1})\right) \\
&\quad + \boldsymbol{w}_m^{\mu-1} \cdot \left(\frac{\eta}{N v_{soft}} (\delta_{l\sigma^{\mu}} - \Phi_l)(\boldsymbol{\xi}^{\mu} - \boldsymbol{w}_l{}^{\mu-1})\right) \\
&\quad + \left(\frac{\eta}{N v_{soft}} (\delta_{m\sigma^{\mu}} - \Phi_m)(\boldsymbol{\xi}^{\mu} - \boldsymbol{w}_m{}^{\mu-1})\right) \\
&\quad \cdot \left(\frac{\eta}{N v_{soft}} (\delta_{l\sigma^{\mu}} - \Phi_l)(\boldsymbol{\xi}^{\mu} - \boldsymbol{w}_l{}^{\mu-1})\right) \\
&= Q_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} \boldsymbol{w}_l{}^{\mu-1} \cdot \left(\delta_{m\sigma^{\mu}} \boldsymbol{\xi}^{\mu} - \delta_{m\sigma^{\mu}} \boldsymbol{w}_m{}^{\mu-1} - \Phi_m \boldsymbol{\xi}^{\mu} + \Phi_m \boldsymbol{w}_m{}^{\mu-1}\right) \\
&\quad + \frac{\eta}{N v_{soft}} \boldsymbol{w}_m^{\mu-1} \cdot \left(\delta_{l\sigma^{\mu}} \boldsymbol{\xi}^{\mu} - \delta_{l\sigma^{\mu}} \boldsymbol{w}_l{}^{\mu-1} - \Phi_l \boldsymbol{\xi}^{\mu} + \Phi_l \boldsymbol{w}_l{}^{\mu-1}\right) \\
&\quad + \left(\frac{\eta}{N v_{soft}}\right)^2 \left(\delta_{m\sigma^{\mu}} \boldsymbol{\xi}^{\mu} - \delta_{m\sigma^{\mu}} \boldsymbol{w}_m{}^{\mu-1} - \Phi_m \boldsymbol{\xi}^{\mu} + \Phi_m \boldsymbol{w}_m{}^{\mu-1}\right) \\
&\quad \cdot \left(\delta_{l\sigma^{\mu}} \boldsymbol{\xi}^{\mu} - \delta_{l\sigma^{\mu}} \boldsymbol{w}_l{}^{\mu-1} - \Phi_l \boldsymbol{\xi}^{\mu} + \Phi_l \boldsymbol{w}_l{}^{\mu-1}\right) \\
&= Q_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} \Big(\delta_{m\sigma^{\mu}} h_l^{\mu} - \delta_{m\sigma^{\mu}} Q_{lm}^{\mu-1} - \Phi_m h_l^{\mu}
\end{aligned}$$

$$
\begin{aligned}
&+\Phi_m Q_{lm}^{\mu-1} + \delta_{l\sigma^\mu} h_m^\mu - \delta_{l\sigma^\mu} Q_{lm}^{\mu-1} - \Phi_l h_m^\mu + \Phi_l Q_{lm}^{\mu-1}\Big) \\
&+\left(\frac{\eta}{N v_{soft}}\right)^2 \Big(\delta_{l\sigma^\mu}\delta_{m\sigma^\mu}(\xi^\mu)^2 - \delta_{l\sigma^\mu}\delta_{m\sigma^\mu} w_l{}^{\mu-1}\xi^\mu - \delta_{m\sigma^\mu}\Phi_l(\xi^\mu)^2 \\
&+\delta_{m\sigma^\mu} w_l{}^{\mu-1}\Phi_l\xi^\mu - \delta_{l\sigma^\mu}\delta_{m\sigma^\mu} w_m{}^{\mu-1}\xi^\mu + \delta_{l\sigma^\mu}\delta_{m\sigma^\mu} w_l{}^{\mu-1} w_m{}^{\mu-1} \\
&+\delta_{m\sigma^\mu} w_m{}^{\mu-1}\Phi_l\xi^\mu - \delta_{m\sigma^\mu} w_l{}^{\mu-1} w_m{}^{\mu-1}\Phi_l - \delta_{l\sigma^\mu}\Phi_m(\xi^\mu)^2 \\
&+\delta_{l\sigma^\mu} w_l{}^{\mu-1}\Phi_m\xi^\mu + \Phi_l\Phi_m(\xi^\mu)^2 - \Phi_l\Phi_m w_l{}^{\mu-1}\xi^\mu \\
&+\delta_{l\sigma^\mu} w_m{}^{\mu-1}\Phi_m\xi^\mu - \delta_{l\sigma^\mu} w_l{}^{\mu-1} w_m{}^{\mu-1}\Phi_m - w_m{}^{\mu-1}\Phi_l\Phi_m\xi^\mu \\
&+w_l{}^{\mu-1} w_m{}^{\mu-1}\Phi_l\Phi_m\Big) \\
&= Q_{lm}^{\mu-1} + Q_{lm}^{\mu-1}\frac{\eta}{N v_{soft}}\Big(\Phi_m + \Phi_l - \delta_{m\sigma^\mu} - \delta_{l\sigma^\mu}\Big) \\
&+ h_l^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{m\sigma^\mu} - \Phi_m\Big) + h_m^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{l\sigma^\mu} - \Phi_l\Big) \\
&+\left(\frac{\eta}{v_{soft}}\right)^2 \Big((\xi^\mu)^2\Big(\delta_{l\sigma^\mu}\delta_{m\sigma^\mu} - \delta_{m\sigma^\mu}\Phi_l - \delta_{l\sigma^\mu}\Phi_m + \Phi_l\Phi_m\Big) \\
&+\delta_{l\sigma^\mu}\delta_{m\sigma^\mu}\Big(Q_{lm} - h_l - h_m\Big) + \delta_{m\sigma^\mu}\Big(h_l\Phi_l + h_m\Phi_l - Q_{lm}\Phi_l\Big) \\
&+\delta_{l\sigma^\mu}\Big(h_l\Phi_m + h_m\Phi_m - Q_{lm}\Phi_m\Big) + \Phi_l\Phi_m\Big(Q_{lm} - h_l - h_m\Big)\Big) \\
&= Q_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} Q_{lm}^{\mu-1}\Big(\Phi_m + \Phi_l - \delta_{m\sigma^\mu} - \delta_{l\sigma^\mu}\Big) \\
&+ h_l^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{m\sigma^\mu} - \Phi_m\Big) + h_m^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{l\sigma^\mu} - \Phi_l\Big) \\
&+\left(\frac{\eta}{v_{soft}}\right)^2 \Big((\xi^\mu)^2\Big(\delta_{l\sigma^\mu}\delta_{m\sigma^\mu} - \delta_{m\sigma^\mu}\Phi_l - \delta_{l\sigma^\mu}\Phi_m + \Phi_l\Phi_m\Big) \\
&+\delta_{l\sigma^\mu}\delta_{m\sigma^\mu}\Big(Q_{lm} - h_l - h_m\Big) + \delta_{m\sigma^\mu}\Big(h_l\Phi_l + h_m\Phi_l - Q_{lm}\Phi_l\Big) \\
&+\delta_{l\sigma^\mu}\Big(h_l\Phi_m + h_m\Phi_m - Q_{lm}\Phi_m\Big) + \Phi_l\Phi_m\Big(Q_{lm} - h_l - h_m\Big)\Big) \\
&\approx Q_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}} Q_{lm}^{\mu-1}\Big(\Phi_m + \Phi_l - \delta_{m\sigma^\mu} - \delta_{l\sigma^\mu}\Big) \\
&+ h_l^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{m\sigma^\mu} - \Phi_m\Big) + h_m^\mu \frac{\eta}{N v_{soft}}\Big(\delta_{l\sigma^\mu} - \Phi_l\Big) \\
&+\left(\frac{\eta}{N v_{soft}}\right)^2 (\xi^\mu)^2\Big(\delta_{l\sigma^\mu}\delta_{m\sigma^\mu} - \delta_{m\sigma^\mu}\Phi_l - \delta_{l\sigma^\mu}\Phi_m + \Phi_l\Phi_m\Big) \qquad (36)
\end{aligned}
$$

In the last step we neglect the terms of $O(\frac{1}{N^2})$ in (36), note that $(\xi^\mu)^2$ is in the order of $O(N)$, see equation (22), so this term remains.

B.2 Differential equations

From equation (35) and (36) follows:

$$\frac{R_{lm}^{\mu} - R_{lm}^{\mu-1}}{1/N} = \frac{\eta}{v_{soft}}\left(\delta_{l\sigma^\mu} b_m^\mu - \delta_{l\sigma^\mu} R_{lm}^{\mu-1} - \Phi_l b_m^\mu + \Phi_l R_{lm}^{\mu-1}\right) \tag{37}$$

$$\begin{aligned}\frac{Q_{lm}^{\mu} - Q_{lm}^{\mu-1}}{1/N} &= \frac{\eta}{v_{soft}} Q_{lm}^{\mu-1}\left(\Phi_m + \Phi_l - \delta_{m\sigma^\mu} - \delta_{l\sigma^\mu}\right) \\ &+ h_l^\mu \frac{\eta}{v_{soft}}\left(\delta_{m\sigma^\mu} - \Phi_m\right) + h_m^\mu \frac{\eta}{v_{soft}}\left(\delta_{l\sigma^\mu} - \Phi_l\right) \\ &+ \frac{1}{N}\left(\frac{\eta}{v_{soft}}\right)^2 (\xi^\mu)^2\left(\delta_{l\sigma^\mu}\delta_{m\sigma^\mu} - \delta_{m\sigma^\mu}\Phi_l - \delta_{l\sigma^\mu}\Phi_m + \Phi_l\Phi_m\right)\end{aligned} \tag{38}$$

We are interested in the mean values of these characteristic quantities, therefore performing averages over the sequence of data. Since the data samples are independent of all previous data samples, the system, including R_{lm} and Q_{lm}, is independent of data sample ξ^μ. Besides this we define:

$$\alpha = \mu/N \tag{39}$$

When considering the thermodynamic limit ($N \to \infty$), α can be considered as a continuous variable with $\Delta\alpha = 1/N$. Using equation (39) and exploiting the independence we get:

$$\begin{aligned}\frac{dR_{lm}}{d\alpha} &= \frac{\eta}{v_{soft}}\Big(< \delta_{l\sigma} b_m > - < \delta_{l\sigma} > R_{lm} - < \Phi_l b_m > + < \Phi_l > R_{lm}\Big) \\ &\quad \left[\begin{array}{l} \because < \delta_{l\sigma} > \;\; = \sum_{\sigma=\pm1} p_\sigma < \delta_{l\sigma} >_\sigma = p_l \\ \because < \delta_{l\sigma} b_m > = \sum_{\sigma=\pm1} p_\sigma < \delta_{l\sigma} b_m >_\sigma = p_l < b_m >_l = p_l \delta_{lm} \lambda \end{array}\right] \\ &= \frac{\eta}{v_{soft}}\Big(p_l(\delta_{lm}\lambda - R_{lm}) - < \Phi_l b_m > + < \Phi_l > R_{lm}\Big)\end{aligned} \tag{40}$$

$$\begin{aligned}\frac{dQ_{lm}}{d\alpha} &= \frac{\eta}{v_{soft}} Q_{lm}\Big(< \Phi_m > + < \Phi_l > - < \delta_{m\sigma} > - < \delta_{l\sigma} > \Big) \\ &+ \frac{\eta}{v_{soft}}\Big(< h_l \delta_{m\sigma} > - < h_l \Phi_m > + < h_m \delta_{l\sigma} > - < h_m \Phi_l > \Big) \\ &+ \frac{1}{N}\left(\frac{\eta}{v_{soft}}\right)^2\Big(< \delta_{l\sigma^\mu}\delta_{m\sigma^\mu}\xi^2 > - < \delta_{m\sigma}\Phi_l\xi^2 > - < \delta_{l\sigma}\Phi_m\xi^2 > + < \Phi_l\Phi_m\xi^2 > \Big) \\ &\quad \left[\begin{array}{l} < X\xi^2 > \;\; \approx \sum_{\sigma=\pm1} p_\sigma << \xi^2 >_\sigma X >_\sigma = \sum_{\sigma=\pm1} p_\sigma v_\sigma N < X >_\sigma \\ \because < \delta_{m\sigma} X > = \sum_{\sigma=\pm1} p_\sigma < \delta_{m\sigma} X >_\sigma = p_m < X >_m \\ \text{and see equations (22, 27)} \end{array}\right] \\ &\approx \frac{\eta}{v_{soft}} Q_{lm}\Big(< \Phi_m > + < \Phi_l > - p_m - p_l\Big) \\ &+ \frac{\eta}{v_{soft}}\Big(p_m\lambda R_{lm} - < h_l \Phi_m > + p_l \lambda R_{ml} - < h_m \Phi_l > \Big)\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{N}\left(\frac{\eta}{v_{soft}}\right)^2\Big(Np_lv_l\delta_{lm}-p_mv_mN<\Phi_l>_m\\
&-p_lv_lN<\Phi_m>_l+\sum_{\sigma=\pm1}p_\sigma v_\sigma N<\Phi_l\Phi_m>_\sigma\Big)\\
&=\frac{\eta}{v_{soft}}Q_{lm}\Big(<\Phi_m>+<\Phi_l>-p_m-p_l\Big)\\
&+\frac{\eta}{v_{soft}}\Big(p_m\lambda R_{lm}-<h_l\Phi_m>+p_l\lambda R_{ml}-<h_m\Phi_l>\Big)\\
&+\left(\frac{\eta}{v_{soft}}\right)^2\Big(\Big(\sum_{\sigma=\pm1}p_\sigma v_\sigma<\Phi_l\Phi_m>_\sigma\Big)+\delta_{lm}p_lv_l-p_mv_m<\Phi_l>_m-p_lv_l<\Phi_m>_l\Big)\\
&\approx\frac{\eta}{v_{soft}}Q_{lm}\Big(<\Phi_m>+<\Phi_l>-p_m-p_l\Big)\\
&+\frac{\eta}{v_{soft}}\Big(p_m\lambda R_{lm}-<h_l\Phi_m>+p_l\lambda R_{ml}-<h_m\Phi_l>\Big)
\end{aligned}
\tag{41}
$$

In the final step we neglect the terms of $O(\eta^2)$, which is correct in the limit $\eta \to 0$. To compute the remaining averages in the differential equations above (40,41) let us look at the function Φ_σ.

$$
\begin{aligned}
\Phi_\sigma &= \Phi\big(\frac{d_{-\sigma}-d_\sigma}{2cv_{soft}}\big)\\
&=\Phi\Big(\frac{1}{2cv_{soft}}\big((\xi-w_{-\sigma})^2-(\xi-w_\sigma)^2\big)\Big)\\
&=\Phi\Big(\frac{1}{2cv_{soft}}\big(2\xi\cdot w_\sigma-2\xi\cdot w_{-\sigma}+w_{-\sigma}^2-w_\sigma^2\big)\Big)\\
&=\Phi\Big(\frac{1}{2cv_{soft}}\big(2h_\sigma-2h_{-\sigma}+Q_{-\sigma-\sigma}-Q_{\sigma\sigma}\big)\Big)\\
&=\Phi\Big(\big(\frac{1}{cv_{soft}},\frac{-1}{cv_{soft}},0,0\big)\cdot(h_{+\sigma},h_{-\sigma},b_{+\sigma},b_{-\sigma})+\frac{Q_{-\sigma-\sigma}-Q_{\sigma\sigma}}{2cv_{soft}}\Big)
\end{aligned}
\tag{42}
$$

Hence we have,

$$
\Phi_\sigma=\Phi\big(\alpha_\sigma\cdot x-\beta_\sigma\big) \tag{43}
$$

Where, $\alpha_\sigma=\big(\frac{\sigma}{cv_{soft}},\frac{-\sigma}{cv_{soft}},0,0\big)$ and $\beta_\sigma=-\big(\frac{Q_{-\sigma,-\sigma}-Q_{\sigma,\sigma}}{2cv_{soft}}\big)$.

Following from equation (20), we can calculate $<\Phi\big(\alpha_\sigma\cdot x-\beta_\sigma\big)>$ and $<(x)_n\Phi\big(\alpha_\sigma\cdot x-\beta_\sigma\big)>$, where $(x)_n$ is the n^{th} component of x; $n\in\{1,2,3,4\}$, see equation (24), from their conditional averages $<\ldots>_\sigma$.

B.3 Averaging Φ

As can be seen in equation (43), we encounter Φ-functions of the following generic form:

$$\Phi_s = \Phi(\alpha_s \cdot x - \beta_s)$$

$$\text{with } \Phi(x) = \int_{-\infty}^{x} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz \tag{44}$$

Now consider averages $< (x)_n \Phi_s >_k$ and $< \Phi_s >_k$:

$$\begin{aligned} < (x)_n \Phi_s >_k &= \frac{1}{(2\pi)^{\frac{4}{2}} (\det(C_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} (x)_n \Phi(\alpha_s \cdot x - \beta_s) \\ &\quad \exp\Big(-\frac{1}{2}(x-\mu_k)^T C_k^{-1} (x-\mu_k)\Big) dx \\ &= \frac{1}{(2\pi)^2 (\det(C_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} (x' + \mu_k)_n \Phi(\alpha_s \cdot x' + \alpha_s \cdot \mu_k - \beta_s) \\ &\quad \exp\Big(-\frac{1}{2} x'^T C_k^{-1} x'\Big) dx' \end{aligned} \tag{45}$$

with $x' = x - \mu_k$. Note that this shift does not change the integral since the limits are $-\infty$ and ∞ in all $\mathbb{R}^4$ dimensions. Now decompose C_k into $C_k = C_k^{\frac{1}{2}} C_k^{\frac{1}{2}}$. This is possible since C_k is a covariance matrix and therefore positive semidefinitive, hence $C_k^{\frac{1}{2}}$ exists. Now let $x' = C_k^{\frac{1}{2}} y$, resulting in $dx' = \det(C_k^{\frac{1}{2}}) dy = (\det(C_k))^{\frac{1}{2}} dy$:

$$\begin{aligned} < (x)_n \Phi_s >_k &= \frac{1}{(2\pi)^2 (\det(C_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} (C_k^{\frac{1}{2}} y + \mu_k)_n \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \\ &\quad \exp\Big(-\frac{1}{2}(C_k^{\frac{1}{2}} y)^T C_k^{-1} C_k^{\frac{1}{2}} y\Big) (\det(C_k))^{\frac{1}{2}} dy \\ &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (C_k^{\frac{1}{2}} y + \mu_k)_n \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \\ &\quad \exp\Big(-\frac{1}{2} y^2\Big) dy \\ &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (C_k^{\frac{1}{2}} y)_n \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2} y^2\Big) dy \\ &+ \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (\mu_k)_n \Phi(\alpha_s \cdot C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2} y^2\Big) dy \\ &= I + \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (\mu_k)_n \Phi(\alpha_s \cdot x' + \alpha_s \cdot \mu_k - \beta_s) \\ &\quad \exp\Big(-\frac{1}{2} (C_k^{-\frac{1}{2}} x')^2\Big) (\det C_k)^{-\frac{1}{2}} dx' \\ &= I + (\mu_k)_n \frac{1}{(2\pi)^2 (\det C_k)^{\frac{1}{2}}} \int_{\mathbb{R}^4} \Phi(\alpha_s \cdot (x - \mu_k) + \alpha_s \cdot \mu_k - \beta_s) \end{aligned}$$

$$
\begin{aligned}
&\exp\Big(-\frac{1}{2}(\boldsymbol{x}-\mu_k)\boldsymbol{C}_k^{-\frac{1}{2}}(\boldsymbol{x}-\mu_k)\Big)d\boldsymbol{x} \\
&= I + (\mu_k)_n \frac{1}{(2\pi)^2(\det \boldsymbol{C}_k)^{\frac{1}{2}}} \int_{\mathbb{R}^4} \Phi(\boldsymbol{\alpha}_s \cdot \boldsymbol{x} - \beta_s) \\
&\quad \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\mu_k)\boldsymbol{C}_k^{-\frac{1}{2}}(\boldsymbol{x}-\mu_k)\Big)d\boldsymbol{x} \\
&= I + (\mu_k)_n < \Phi_s >_k
\end{aligned} \tag{46}
$$

Where

$$
\begin{aligned}
I &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} (\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y})_n \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2}\boldsymbol{y}^2\Big)d\boldsymbol{y} \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} \sum_{j=1}^{4} \Big((\boldsymbol{C}_k^{\frac{1}{2}})_{nj}(y)_j\Big) \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2}\boldsymbol{y}^2\Big)d\boldsymbol{y} \\
&= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left(\int_{\mathbb{R}} (\boldsymbol{C}_k^{\frac{1}{2}})_{nj}(y)_j \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2}(y)_j^2\Big) d(y)_j \right)
\end{aligned} \tag{47}
$$

Now define, for $j \in \{1,2,3,4\}$:

$$
I_j = \int_{\mathbb{R}} (\boldsymbol{C}_k^{\frac{1}{2}})_{nj}(y)_j \Phi(\boldsymbol{\alpha}_s \cdot \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2}(y)_j^2\Big) d(y)_j \tag{48}
$$

So one can write:

$$
I = \frac{1}{(2\pi)^2} \sum_{j=1}^{4} I_j \tag{49}
$$

Now apply partial integration to I_j. The rule of partial integration says: $\int f(x)g'(x)dx = f(x)g(x) - \int g(x)f'(x)dx$. Applied to I_j this gives:

$$
\begin{aligned}
f((y)_j) &= \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \\
g'((y)_j) &= (\boldsymbol{C}_k^{\frac{1}{2}})_{nj}(y)_j \exp\big(-\frac{1}{2}(y)_j^2\big) \\
g((y)_j) &= -(\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\big(-\frac{1}{2}(y)_j^2\big) \\
f'((y)_j) &= \frac{\partial}{\partial (y)_j} \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s)
\end{aligned} \tag{50}
$$

Applying the partial integration yields:

$$
I_j = \underbrace{\left[-(\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\big(-\frac{1}{2}(y)_j^2\big) \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \mu_k - \beta_s) \right]_{-\infty}^{\infty}}_{=0}
$$

$$+ \int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} \exp\big(-\frac{1}{2}(y)_j^2\big) \frac{\partial}{\partial (y)_j} \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) d(y)_j$$
$$= \int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} \exp\big(-\frac{1}{2}(y)_j^2\big) \frac{\partial}{\partial (y)_j} \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) d(v)_j \quad (51)$$

Filling this in into equation (49) gives:

$$I = \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left(\int_{\mathbb{R}} (C_k^{\frac{1}{2}})_{nj} \exp\big(-\frac{1}{2}(y)_j^2\big) \frac{\partial}{\partial (y)_j} \Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) d(y)_j \right)$$
$$= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left((C_k^{\frac{1}{2}})_{nj} \right) \int_{\mathbb{R}^4} \frac{\partial}{\partial (y)_j} \Big(\Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \Big) \exp\Big(-\frac{1}{2} y^2\Big) dy \quad (52)$$

Now consider:

$$\frac{\partial \Big(\Phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \Big)}{\partial (y)_j} = \phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \frac{\partial}{\partial (y)_j} \Big(\alpha_s C_k^{\frac{1}{2}} y \Big)$$
$$= \sum_{i=1}^{4} (\alpha_s)_i (C_k^{\frac{1}{2}})_{ij} \phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \quad (53)$$

Where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the standard normal probability density function and in the last step is used that for the components $(y)_i$ with $i \neq j$ the derivative to $(y)_j$ is zero, resulting in: $\frac{\partial}{\partial (y)_j} \Big(\alpha_s C_k^{\frac{1}{2}} y \Big) = \sum_{i=1}^{4} (\alpha_s)_i (C_k^{\frac{1}{2}})_{ij}$. Hence,

$$I = \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \Big((C_k^{\frac{1}{2}})_{nj} \sum_{i=1}^{4} (\alpha_s)_i (C_k^{\frac{1}{2}})_{ij} \Big)$$
$$\int_{\mathbb{R}^4} \phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2} y^2\Big) dy$$
$$= \frac{1}{(2\pi)^2} (C_k \alpha_s)_n \int_{\mathbb{R}^4} \phi(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s) \exp\Big(-\frac{1}{2} y^2\Big) dy \quad (54)$$

Note that in the last equation it is used that $\sum_{j=1}^{4} \Big((C_k^{\frac{1}{2}})_{nj} \sum_{i=1}^{4} (\alpha_s)_i (C_k^{\frac{1}{2}})_{ij} \Big) = (C_k \alpha_s)_n$, which is true for symmetrical matrices $C_k^{\frac{1}{2}}$. Since C_k is a covariance matrix hence it is symmetrical, see equation (34), and positive definite, there exists at least one decomposition into $C_k^{\frac{1}{2}}$ also being symmetrical.

Note also that $\exp(-\frac{1}{2} y^2) dy$ is a measure which is invariant under rotation of the coordinate axes. Now rotate the system in such a way that one of the axes, say $\tilde{y}$, is aligned with the vector $C_k^{\frac{1}{2}} \alpha_s$. Using that $\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(-\frac{1}{2} z^2) dz = 1$, the remaining three coordinates can be integrated over, yielding:

$$I = \frac{1}{\sqrt{2\pi}}(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n \int_{\mathbb{R}} \phi(\|\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{\alpha}_s\|\tilde{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s) \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y} \quad (55)$$

Now define:

$$\tilde{\alpha}_{s,k} = \|\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{\alpha}_s\| = \sqrt{\boldsymbol{\alpha}_s\boldsymbol{C}_k\boldsymbol{\alpha}_s} \quad \text{and} \quad \tilde{\beta}_{s,k} = \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s \quad (56)$$

and rename $z = \tilde{\alpha}_{s,k}\tilde{y}$, so $d\tilde{y} = \frac{1}{\tilde{\alpha}_{s,k}} dz$, to obtain:

$$\begin{aligned}
I &= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}} \int_{\mathbb{R}} \phi(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}) \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y} \\
&= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \int_{\mathbb{R}} \phi(z + \tilde{\beta}_{s,k}) \exp\Big(-\frac{1}{2}(\frac{z}{\tilde{\alpha}_{s,k}})^2\Big) dz \\
&= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(z + \tilde{\beta}_{s,k})^2}{2}) \exp\Big(-\frac{1}{2}(\frac{z}{\tilde{\alpha}_{s,k}})^2\Big) dz \\
&= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}z^2(1 + \frac{1}{\tilde{\alpha}_{s,k}^2}) - z\tilde{\beta}_{s,k} - \frac{1}{2}\tilde{\beta}_{s,k}^2\Big) dz \\
&= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \exp(-\frac{1}{2}\tilde{\beta}_{s,k}^2) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}z^2(1 + \frac{1}{\tilde{\alpha}_{s,k}^2}) - z\tilde{\beta}_{s,k}\Big) dz \quad (57)
\end{aligned}$$

Now use that $\int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}} \exp(-\frac{1}{2}ax^2 + bx) = \frac{1}{\sqrt{a}} \exp(\frac{b^2}{2a})$:

$$\begin{aligned}
I &= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \exp(-\frac{1}{2}\tilde{\beta}_{s,k}^2) \frac{1}{\sqrt{1 + \frac{1}{\tilde{\alpha}_{s,k}^2}}} \exp(\frac{\tilde{\beta}_{s,k}^2}{2(1 + \frac{1}{\tilde{\alpha}_{s,k}^2})}) \\
&= \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \frac{1}{\sqrt{1 + \frac{1}{\tilde{\alpha}_{s,k}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{s,k}^2}{(1 + \tilde{\alpha}_{s,k}^2)}\Big) \quad (58)
\end{aligned}$$

Summarizing equations (47) to (58) gives, from equation (45):

$$< (\boldsymbol{x})_n\Phi_s >_k = \frac{(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \frac{1}{\sqrt{1 + \frac{1}{\tilde{\alpha}_{s,k}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{s,k}^2}{(1 + \tilde{\alpha}_{s,k}^2)}\Big) + (\boldsymbol{\mu}_k)_n < \Phi_s >_k \quad (59)$$

Next consider the average $< \Phi_s >_k$. Similar calculations as for $< (\boldsymbol{x})_n\Phi_s >_k$ give:

$$\begin{aligned}
< \Phi_s >_k = \frac{1}{(2\pi)^{\frac{4}{2}}(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} &\Phi(\boldsymbol{\alpha}_s \cdot \boldsymbol{x} - \beta_s) \\
&\exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{C}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\Big) d\boldsymbol{x}
\end{aligned}$$

$$= \frac{1}{(2\pi)^2(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathrm{R}^4} \Phi(\boldsymbol{\alpha}_s \cdot \boldsymbol{x}' + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s)$$
$$\exp\Big(-\frac{1}{2}\boldsymbol{x}'^T \boldsymbol{C}_k^{-1}\boldsymbol{x}'\Big)d\boldsymbol{x}'$$
$$= \frac{1}{(2\pi)^2(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathrm{R}^4} \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} y + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s)$$
$$\exp\Big(-\frac{1}{2}(\boldsymbol{C}_k^{\frac{1}{2}} y)^T \boldsymbol{C}_k^{-1} \boldsymbol{C}_k^{\frac{1}{2}} y\Big)(\det(\boldsymbol{C}_k))^{\frac{1}{2}} dy$$
$$= \frac{1}{(2\pi)^2} \int_{\mathrm{R}^4} \Phi(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} y + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s) \exp\Big(-\frac{1}{2}y^2\Big)dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{\mathrm{R}} \Phi(\|\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\|\tilde{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s) \exp\Big(-\frac{1}{2}\tilde{y}^2\Big)d\tilde{y}$$
$$= \frac{1}{\sqrt{2\pi}} \int_{\mathrm{R}} \Phi(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}) \exp\Big(-\frac{1}{2}\tilde{y}^2\Big)d\tilde{y} \tag{60}$$

Now apply $\int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}} \exp(-\frac{1}{2}ax^2 + bx)\Phi(cx + d) = \frac{1}{\sqrt{a}} \exp\Big(\frac{b^2}{2a}\Big)\Phi\Big(\frac{bc+ad}{\sqrt{a^2+ac^2}}\Big)$, which holds for $(a > 0)$:

$$< \Phi_s >_k = \Phi\Big(\frac{\tilde{\beta}_{s,k}}{\sqrt{1 + \tilde{\alpha}_{s,k}^2}}\Big) \tag{61}$$

Hence the required averages are as follows:

$$< \Phi_s >_k = \Phi\Big(\frac{\tilde{\beta}_{s,k}}{\sqrt{1 + \tilde{\alpha}_{s,k}^2}}\Big)$$
$$< (x)_n \Phi_s >_k = \frac{(\boldsymbol{C}_k \boldsymbol{\alpha}_s)_n}{\sqrt{2\pi}\tilde{\alpha}_{s,k}} \frac{1}{\sqrt{1 + \frac{1}{\tilde{\alpha}_{s,k}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{s,k}^2}{(1 + \tilde{\alpha}_{s,k}^2)}\Big)$$
$$+ (\boldsymbol{\mu}_k)_n \Phi\Big(\frac{\tilde{\beta}_{s,k}}{\sqrt{1 + \tilde{\alpha}_{s,k}^2}}\Big) \tag{62}$$

Where,

$$\tilde{\alpha}_{s,k} = \sqrt{\boldsymbol{\alpha}_s \boldsymbol{C}_k \boldsymbol{\alpha}_s} \quad \text{and} \quad \tilde{\beta}_{s,k} = \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s \tag{63}$$

B.4 Final form of the differential equations

Filling in the conditional averages of equation (62) into equations (40) and(41), yields the following system of coupled differential equations:

$$\begin{aligned}
\frac{dR_{lm}}{d\alpha} &= \frac{\eta}{v_{soft}}\Big(p_l(\delta_{lm}\lambda - R_{lm}) - \sum_{\sigma=\pm 1} p_\sigma < b_m \Phi\big(\boldsymbol{\alpha}_l \cdot \boldsymbol{x} - \beta_l\big) >_\sigma \\
&\quad + R_{lm} \sum_{\sigma=\pm 1} p_\sigma < \Phi\big(\boldsymbol{\alpha}_l \cdot \boldsymbol{x} - \beta_l\big) >_\sigma \Big) \\
&= \frac{\eta}{v_{soft}}\Bigg(p_l(\delta_{lm}\lambda - R_{lm}) \\
&\quad - \Big[\sum_{\sigma=\pm 1} p_\sigma \Bigg(\frac{(\boldsymbol{C}_\sigma \boldsymbol{\alpha}_l)_{n_{bm}}}{\sqrt{2\pi}\tilde{\alpha}_{l,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}^2_{l,\sigma}}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}^2_{l,\sigma}}{(1+\tilde{\alpha}^2_{l,\sigma})}\Big) \\
&\quad + (\boldsymbol{\mu}_\sigma)_{n_{bm}} \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{l,\sigma}}}\Big)\Bigg)\Big] + R_{lm} \sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{l,\sigma}}}\Big)\Bigg)
\end{aligned} \tag{64}$$

$$\begin{aligned}
\frac{dQ_{lm}}{d\alpha} &= \frac{\eta}{v_{soft}} Q_{lm}\Big(\sum_{\sigma=\pm 1} p_\sigma < \Phi\big(\boldsymbol{\alpha}_m \cdot \boldsymbol{x} - \beta_m\big) >_\sigma + \sum_{\sigma=\pm 1} p_\sigma < \Phi\big(\boldsymbol{\alpha}_l \cdot \boldsymbol{x} - \beta_l\big) >_\sigma - p_m - p_l\Big) \\
&\quad + \frac{\eta}{v_{soft}}\Big(p_m \lambda R_{lm} - \sum_{\sigma=\pm 1} p_\sigma < h_l \Phi\big(\boldsymbol{\alpha}_m \cdot \boldsymbol{x} - \beta_m\big) >_\sigma \\
&\quad + p_l \lambda R_{ml} - \sum_{\sigma=\pm 1} p_\sigma < h_m \Phi\big(\boldsymbol{\alpha}_l \cdot \boldsymbol{x} - \beta_l\big) >_\sigma \Big) \\
&= \frac{\eta}{v_{soft}} Q_{lm}\Big(\sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{m,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{m,\sigma}}}\Big) + \sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{l,\sigma}}}\Big) - p_m - p_l\Big) \\
&\quad + \frac{\eta}{v_{soft}}\Bigg(p_m \lambda R_{lm} + p_l \lambda R_{ml} \\
&\quad - \Bigg[\sum_{\sigma=\pm 1} p_\sigma \Bigg(\frac{(\boldsymbol{C}_\sigma \boldsymbol{\alpha}_m)_{n_{hl}}}{\sqrt{2\pi}\tilde{\alpha}_{m,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}^2_{m,\sigma}}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}^2_{m,\sigma}}{(1+\tilde{\alpha}^2_{m,\sigma})}\Big) + (\boldsymbol{\mu}_\sigma)_{n_{hl}} \Phi\Big(\frac{\tilde{\beta}_{m,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{m,\sigma}}}\Big)\Bigg)\Bigg] \\
&\quad - \Bigg[\sum_{\sigma=\pm 1} p_\sigma \Bigg(\frac{(\boldsymbol{C}_\sigma \boldsymbol{\alpha}_l)_{n_{hm}}}{\sqrt{2\pi}\tilde{\alpha}_{l,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}^2_{l,\sigma}}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}^2_{l,\sigma}}{(1+\tilde{\alpha}^2_{l,\sigma})}\Big) + (\boldsymbol{\mu}_\sigma)_{n_{hm}} \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}^2_{l,\sigma}}}\Big)\Bigg)\Bigg]\Bigg)
\end{aligned} \tag{65}$$

Where,

$$n_{bk} = \begin{cases} 3 \text{ if } k = 1 \\ 4 \text{ if } k = -1 \end{cases}$$

$$n_{hk} = \begin{cases} 1 \text{ if } k = 1 \\ 2 \text{ if } k = -1 \end{cases}$$

$$\tilde{\alpha}_{s,k} = \sqrt{\alpha_s C_k \alpha_s}\ ,\ \tilde{\beta}_{s,k} = \alpha_s \cdot \mu_k - \beta_s$$
$$\alpha_\sigma = \left(\tfrac{\sigma}{cv_{soft}}, \tfrac{-\sigma}{cv_{soft}}, 0, 0\right)\ ,\ \beta_\sigma = -\left(\tfrac{Q_{-\sigma,-\sigma} - Q_{\sigma,\sigma}}{2cv_{soft}}\right) \tag{66}$$

In equations (64) and (65) one can see that the ODE's are linear in η. The learning rate η is therefore, in the limit $\eta \to 0$, no more than a scaling factor, which can be taken out by rescaling α to $\tilde{\alpha} = \alpha\eta$:

$$\begin{aligned}
\frac{dR_{lm}}{d\tilde{\alpha}} &= \frac{dR_{lm}}{d\alpha\eta} \\
&= \frac{1}{v_{soft}}\Bigg(p_l(\delta_{lm}\lambda - R_{lm}) \\
&\quad -\Big[\sum_{\sigma=\pm 1} p_\sigma\Bigg(\frac{(C_\sigma \alpha_l)_{n_{bm}}}{\sqrt{2\pi}\tilde{\alpha}_{l,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}_{l,\sigma}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{l,\sigma}^2}{(1+\tilde{\alpha}_{l,\sigma}^2)}\Big) \\
&\quad +(\mu_\sigma)_{n_{bm}} \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}_{l,\sigma}^2}}\Big)\Bigg)\Big] + R_{lm}\sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}_{l,\sigma}^2}}\Big)\Bigg)
\end{aligned} \tag{67}$$

$$\begin{aligned}
\frac{dQ_{lm}}{d\tilde{\alpha}} &= \frac{dQ_{lm}}{d\alpha\eta} \\
&= \frac{1}{v_{soft}} Q_{lm}\Bigg(\sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{m,\sigma}}{\sqrt{1+\tilde{\alpha}_{m,\sigma}^2}}\Big) + \sum_{\sigma=\pm 1} p_\sigma \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}_{l,\sigma}^2}}\Big) - p_m - p_l\Bigg) \\
&\quad + \frac{1}{v_{soft}}\Bigg(p_m \lambda R_{lm} + p_l \lambda R_{ml} \\
&\quad -\Bigg[\sum_{\sigma=\pm 1} p_\sigma\Bigg(\frac{(C_\sigma \alpha_m)_{n_{hl}}}{\sqrt{2\pi}\tilde{\alpha}_{m,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}_{m,\sigma}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{m,\sigma}^2}{(1+\tilde{\alpha}_{m,\sigma}^2)}\Big) + (\mu_\sigma)_{n_{hl}} \Phi\Big(\frac{\tilde{\beta}_{m,\sigma}}{\sqrt{1+\tilde{\alpha}_{m,\sigma}^2}}\Big)\Bigg)\Bigg] \\
&\quad -\Bigg[\sum_{\sigma=\pm 1} p_\sigma\Bigg(\frac{(C_\sigma \alpha_l)_{n_{hm}}}{\sqrt{2\pi}\tilde{\alpha}_{l,\sigma}} \frac{1}{\sqrt{1+\frac{1}{\tilde{\alpha}_{l,\sigma}^2}}} \exp\Big(-\frac{1}{2}\frac{\tilde{\beta}_{l,\sigma}^2}{(1+\tilde{\alpha}_{l,\sigma}^2)}\Big) + (\mu_\sigma)_{n_{hm}} \Phi\Big(\frac{\tilde{\beta}_{l,\sigma}}{\sqrt{1+\tilde{\alpha}_{l,\sigma}^2}}\Big)\Bigg)\Bigg]\Bigg)
\end{aligned} \tag{68}$$

B.5 Original RSLVQ

Now let us look at the original definition of RSLVQ by Obermayer and Seo. The recurrence relations are the same, except we need to replace Φ_l by $\frac{1}{1+\exp(\frac{d_l^\mu - d_{-l}^\mu}{2v_{soft}})}$ or Ω_l for shorthand.

Recurrence relation for R: From equation (35) follows:

$$R_{lm}^\mu = R_{lm}^{\mu-1} + \frac{\eta}{N v_{soft}}\Big(\delta_{l\sigma^\mu} b_m^\mu - \delta_{l\sigma^\mu} R_{lm}^{\mu-1} - \Omega_l b_m^\mu + \Omega_l R_{lm}^{\mu-1}\Big) \tag{69}$$

Recurrence relation for Q: From equation (36) follows:

$$
\begin{aligned}
Q^{\mu}_{lm} \approx\ & Q^{\mu-1}_{lm} + \frac{\eta}{N v_{soft}} Q^{\mu-1}_{lm}\Big(\Omega_m + \Omega_l - \delta_{m\sigma^\mu} - \delta_{l\sigma^\mu}\Big) \\
& + h^{\mu}_{l}\frac{\eta}{N v_{soft}}\Big(\delta_{m\sigma^\mu} - \Omega_m\Big) + h^{\mu}_{m}\frac{\eta}{N v_{soft}}\Big(\delta_{l\sigma^\mu} - \Omega_l\Big) \\
& + \left(\frac{\eta}{N v_{soft}}\right)^2 (\xi^\mu)^2\Big(\delta_{lm} - \delta_{m\sigma^\mu}\Omega_l - \delta_{l\sigma^\mu}\Omega_m + \Omega_l\Omega_m\Big)
\end{aligned}
\tag{70}
$$

Differential equations: From equation (69) and (70) follows, similar to equations (40) and (41):

$$
\frac{dR_{lm}}{d\alpha} = \frac{\eta}{v_{soft}}\Big(p_l(\delta_{lm}\lambda - R_{lm}) - < \Omega_l b_m > + < \Omega_l > R_{lm}\Big) \tag{71}
$$

$$
\begin{aligned}
\frac{dQ_{lm}}{d\alpha} \approx\ & \frac{\eta}{v_{soft}} Q_{lm}\Big(< \Omega_m > + < \Omega_l > - p_m - p_l\Big) \\
& + \frac{\eta}{v_{soft}}\Big(p_m\lambda R_l - < h_l\Omega_m > + p_l\lambda R_{ml} - < h_m\Omega_l > \Big) \\
& + \left(\frac{\eta}{v_{soft}}\right)^2\Big(\Big(\sum_{\sigma=\pm 1} p_\sigma v_\sigma(< \Omega_l\Omega_m >_\sigma + \delta_{lm})\Big) - p_m v_m < \Omega_l >_m - p_l v_l < \Omega_m >_l \Big) \\
\approx\ & \frac{\eta}{v_{soft}} Q_{lm}\Big(< \Omega_m > + < \Omega_l > - p_m - p_l\Big) \\
& + \frac{\eta}{v_{soft}}\Big(p_m\lambda R_l - < h_l\Omega_m > + p_l\lambda R_{ml} - < h_m\Omega_l > \Big)
\end{aligned}
\tag{72}
$$

Note again that in the final step we neglect the terms of $O(\eta^2)$, which is correct in the limit $\eta \to 0$. Similar to equation (42) we can generalize Ω_σ.

$$
\begin{aligned}
\Omega_\sigma &= \frac{1}{1 + \exp\left(\frac{d_\sigma - d_{-\sigma}}{2v_{soft}}\right)} \\
&= \frac{1}{1 + \exp\left(\frac{1}{2v_{soft}}\left((\xi - w_\sigma)^2 - (\xi - w_{-\sigma})^2\right)\right)} \\
&= \frac{1}{1 + \exp\left(\frac{1}{2v_{soft}}\left(2\xi\cdot w_{-\sigma} - 2\xi\cdot w_\sigma + w_\sigma^2 - w_{-\sigma}^2\right)\right)} \\
&= \frac{1}{1 + \exp\left(\frac{1}{2v_{soft}}\left(2h_{-\sigma} - 2h_\sigma + Q_{\sigma\sigma} - Q_{-\sigma-\sigma}\right)\right)} \\
&= \frac{1}{1 + \exp\left(\left(\frac{-1}{v_{soft}}, \frac{1}{v_{soft}}, 0, 0\right)\cdot(h_{+\sigma}, h_{-\sigma}, b_{+\sigma}, b_{-\sigma}) + \frac{Q_{\sigma\sigma} - Q_{-\sigma-\sigma}}{2v_{soft}}\right)} \\
&= \frac{1}{1 + \exp\left(\alpha_\sigma\cdot x + \beta_\sigma\right)}
\end{aligned}
\tag{73}
$$

Where, $\alpha_\sigma = \left(\frac{-\sigma}{v_{soft}}, \frac{\sigma}{v_{soft}}, 0, 0\right)$ and $\beta_\sigma = -\left(\frac{Q_{\sigma\sigma} - Q_{-\sigma-\sigma}}{2v_{soft}}\right)$.

Now consider the averages $< (x)_n \Omega_s >_k$ and $< \Omega_s >_k$:

$$
\begin{aligned}
< (x)_n \Omega_s >_k &= \frac{1}{(2\pi)^{\frac{4}{2}} (\det(C_k))^{\frac{1}{2}}} \int_{\mathrm{R}^4} \frac{(x)_n}{1 + \exp\left(\alpha_s \cdot x + \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2}(x - \mu_k)^T C_k^{-1} (x - \mu_k)\Big) dx \\
&= \frac{1}{(2\pi)^2 (\det(C_k))^{\frac{1}{2}}} \int_{\mathrm{R}^4} \frac{(x' + \mu_k)_n}{1 + \exp\left(\alpha_s \cdot x' + \alpha_s \cdot \mu_k - \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2} x'^T C_k^{-1} x'\Big) dx' \\
&= \frac{1}{(2\pi)^2 (\det(C_k))^{\frac{1}{2}}} \int_{\mathrm{R}^4} \frac{(C_k^{\frac{1}{2}} y + \mu_k)_n}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2}(C_k^{\frac{1}{2}} y)^T C_k^{-1} C_k^{\frac{1}{2}} y\Big) (\det(C_k))^{\frac{1}{2}} dy \\
&= \frac{1}{(2\pi)^2} \int_{\mathrm{R}^4} \frac{(C_k^{\frac{1}{2}} y)_n}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \exp\Big(-\frac{1}{2} y^2\Big) dy \\
&+ \frac{1}{(2\pi)^2} \int_{\mathrm{R}^4} \frac{(\mu_k)_n}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \exp\Big(-\frac{1}{2} y^2\Big) dy \\
&= I + (\mu_k)_n < \Omega_s >_k
\end{aligned} \tag{74}
$$

Where

$$
\begin{aligned}
I &= \frac{1}{(2\pi)^2} \int_{\mathrm{R}^4} \frac{(C_k^{\frac{1}{2}} y)_n}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \exp\Big(-\frac{1}{2} y^2\Big) dy \\
&= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left(\int_{\mathrm{R}} \frac{(C_k^{\frac{1}{2}})_{nj} (y)_j}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \exp\Big(-\frac{1}{2}(y)_j^2\Big) d(y)_j \right) \\
&= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} I_j
\end{aligned} \tag{75}
$$

With

$$
I_j = \int_{\mathrm{R}} \frac{(C_k^{\frac{1}{2}})_{nj} (y)_j}{1 + \exp\left(\alpha_s C_k^{\frac{1}{2}} y + \alpha_s \cdot \mu_k - \beta_s\right)} \exp\Big(-\frac{1}{2}(y)_j^2\Big) d(y)_j \tag{76}
$$

Now apply partial integration to I_j. The rule of partial integration says: $\int f(x)g'(x)dx = f(x)g(x) - \int g(x)f'(x)dx$. Applied to I_j this gives:

$$\begin{aligned} f((\boldsymbol{y})_j) &= \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} \\ g'((\boldsymbol{y})_j) &= (\boldsymbol{C}_k^{\frac{1}{2}})_{nj} (\boldsymbol{y})_j \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \\ g((\boldsymbol{y})_j) &= -(\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \\ f'((\boldsymbol{y})_j) &= \frac{\partial}{\partial(\boldsymbol{y})_j} \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} \end{aligned} \tag{77}$$

Applying the partial integration yields:

$$\begin{aligned} I_j &= \underbrace{\left[-(\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} \right]_{-\infty}^{\infty}}_{=0} \\ &\quad + \int_{\mathbb{R}} (\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \frac{\partial}{\partial(\boldsymbol{y})_j} \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} d(\boldsymbol{y})_j \\ &= \int_{\mathbb{R}} (\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \frac{\partial}{\partial(\boldsymbol{y})_j} \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} d(v)_j \end{aligned} \tag{78}$$

Filling this in into equation (75) gives:

$$\begin{aligned} I &= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left(\int_{\mathbb{R}} (\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \exp\left(-\frac{1}{2}(\boldsymbol{y})_j^2\right) \frac{\partial}{\partial(\boldsymbol{y})_j} \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} d(\boldsymbol{y})_j \right) \\ &= \frac{1}{(2\pi)^2} \sum_{j=1}^{4} \left((\boldsymbol{C}_k^{\frac{1}{2}})_{nj} \right) \int_{\mathbb{R}^4} \frac{\partial}{\partial(\boldsymbol{y})_j} \left(\frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)} \right) \exp\left(-\frac{1}{2}\boldsymbol{y}^2\right) d\boldsymbol{y} \end{aligned} \tag{79}$$

Now consider:

$$\begin{aligned} \frac{\partial \frac{1}{1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)}}{\partial(\boldsymbol{y})_j} &= -\frac{\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)}{\left(1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)\right)^2} \frac{\partial}{\partial(\boldsymbol{y})_j} \left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y}\right) \\ &= -\sum_{i=1}^{4} (\boldsymbol{\alpha}_s)_i (\boldsymbol{C}_k^{\frac{1}{2}})_{ij} \frac{\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)}{\left(1+\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \boldsymbol{\beta}_s\right)\right)^2} \end{aligned} \tag{80}$$

$$I = -\frac{1}{(2\pi)^2}(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n \int_{\mathbb{R}^4} \frac{\exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)}{\left(1 + \exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)\right)^2} \exp\Big(-\frac{1}{2}\boldsymbol{y}^2\Big) d\boldsymbol{y}$$

$$= -\frac{1}{\sqrt{2\pi}}(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n \int_{\mathbf{R}} \frac{\exp\left(\|\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{\alpha}_s\|\tilde{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)}{\left(1 + \exp\left(\|\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{\alpha}_s\|\tilde{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)\right)^2} \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y}$$

$$= -\frac{1}{\sqrt{2\pi}}(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n \int_{\mathbf{R}} \frac{\exp\left(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}\right)}{\left(1 + \exp\left(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}\right)\right)^2} \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y} \tag{81}$$

Filling this in into equation (74) gives:

$$< (\boldsymbol{x})_n \Omega_s >_k = -\frac{1}{\sqrt{2\pi}}(\boldsymbol{C}_k\boldsymbol{\alpha}_s)_n \int_{\mathbf{R}} \frac{\exp\left(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}\right)}{\left(1 + \exp\left(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}\right)\right)^2} \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y} + (\mu_k)_n < \Omega_s >_k \tag{82}$$

Next consider the average $< \Omega_s >_k$:

$$\begin{aligned}
< \Omega_s >_k &= \frac{1}{(2\pi)^{\frac{4}{2}}(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathbb{R}^4} \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_s \cdot \boldsymbol{x} + \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{C}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\Big) d\boldsymbol{x} \\
&= \frac{1}{(2\pi)^2(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathbf{R}^4} \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_s \cdot \boldsymbol{x}' + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2}\boldsymbol{x}'^T \boldsymbol{C}_k^{-1} \boldsymbol{x}'\Big) d\boldsymbol{x}' \\
&= \frac{1}{(2\pi)^2(\det(\boldsymbol{C}_k))^{\frac{1}{2}}} \int_{\mathbf{R}^4} \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)} \\
&\quad \exp\Big(-\frac{1}{2}(\boldsymbol{C}_k^{\frac{1}{2}}\boldsymbol{y})^T \boldsymbol{C}_k^{-1} \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y}\Big) (\det(\boldsymbol{C}_k))^{\frac{1}{2}} d\boldsymbol{y} \\
&= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}} \boldsymbol{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)} \exp\Big(-\frac{1}{2}\boldsymbol{y}^2\Big) d\boldsymbol{y} \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \frac{1}{1 + \exp\left(\|\boldsymbol{\alpha}_s \boldsymbol{C}_k^{\frac{1}{2}}\|\tilde{y} + \boldsymbol{\alpha}_s \cdot \boldsymbol{\mu}_k - \beta_s\right)} \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y} \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \frac{1}{1 + \exp\left(\tilde{\alpha}_{s,k}\tilde{y} + \tilde{\beta}_{s,k}\right)} \exp\Big(-\frac{1}{2}\tilde{y}^2\Big) d\tilde{y}
\end{aligned} \tag{83}$$

where the remaining integration has to be performed numerically.