Stochastic Dynamics of Learning with Momentum in Neural Networks

Wim Wiegerinck\textsuperscript{1}  Andrzej Komoda\textsuperscript{1}  Tom Heskes\textsuperscript{2}

\textsuperscript{1}Department of Medical Physics and Biophysics, University of Nijmegen, Geert Grooteplein Noord 21, 6525 EZ Nijmegen, The Netherlands, e-mail: wimw@mbfys.kun.nl

\textsuperscript{2}Beckman Institute and Department of Physics, University of Illinois, 405 North Mathews Avenue, Urbana, Illinois 61801, U.S.A.

Abstract

We study on-line learning with momentum term for nonlinear learning rules. Through introduction of auxiliary variables, we show that the learning process can be described by a Markov process.

For small learning parameters \(\eta\) and momentum parameters \(\alpha\) close to 1, such that \(\gamma = \eta/(1 - \alpha)^2\) is finite, the time scales for the evolution of the weights and the auxiliary variables are the same. In this case Van Kampen’s expansion can be applied in a straightforward manner. We obtain evolution equations for the average network state and the fluctuations around this average. These evolution equations depend (after rescaling of time and fluctuations) only on \(\gamma\): all combinations \((\eta, \alpha)\) with the same value of \(\gamma\) give rise to similar behaviour.

The case \(\alpha\) constant and \(\eta\) small requires a completely different analysis. There are two different time scales: a fast time scale on which the auxiliary variables equilibrate and a slow time scale for the change of the weights. By projection on the space of slow variables the fast variables can be eliminated. We find that for small learning parameters \(\eta\) and finite momentum parameters \(\alpha\) learning with momentum is equivalent to learning without momentum term with rescaled learning parameter \(\tilde{\eta} = \eta/(1 - \alpha)\).

Simulations with the nonlinear Oja learning rule confirm the theoretical results.

PACS: 87.10.+e, 02.50.-r, 05.40.+j
1 Introduction

1.1 Background

On-line learning stands for learning in artificial neural networks where a weight change takes place each time a training pattern \( x \) is drawn at random from the total training set and is presented to the network. This weight change can be written in the general form

\[
\Delta w(n) \equiv w(n+1) - w(n) = \eta f(w(n), x),
\]

with \( w(n) \) the network state at iteration step \( n \), \( \eta \) the so-called learning parameter, and \( f(\cdot, \cdot) \) the learning rule. Because of the random presentation of patterns \( x \), on-line learning as described by (1) is a stochastic process. The probability to be in a certain network state \( w \) can be shown to obey a master equation. In recent years, theoretical studies of this master equation have provided a better understanding of on-line learning processes [1, 2, 3, 4, 5, 6].

When the last weight change is added to the learning rule (1), the weight change takes the form

\[
\Delta w(n) = \eta f(w(n), x) + \alpha \Delta w(n-1),
\]

with \( \alpha \) the so-called momentum parameter. Equation (2) describes on-line learning with momentum term. The incorporation of this momentum term is frequently applied to backpropagation [7] with the intention to speed up learning with essentially no increase in computational complexity (see e.g. [8] for references). The backpropagation learning rule, like most learning rules in neural network literature, is nonlinear in the weights \( w \). Theoretical studies, however, have been mainly focusing on the linear LMS algorithm with momentum updating [9, 10]. In this paper we will consider the effect of the momentum term on general nonlinear learning rules and ask ourselves whether incorporation of the momentum term leads to an improvement of the performance of on-line learning rules.

1.2 Framework

Equation (2) describes a second-order process. It can be turned into a Markov process through the introduction of the auxiliary variable \( \mu(n) \equiv \Delta w(n-1) \):

\[
\begin{align*}
\Delta w(n) &= \eta f(w(n), x) + \alpha \mu(n) \\
\Delta \mu(n) &= \eta f(w(n), x) + (\alpha - 1) \mu(n).
\end{align*}
\]

With definitions \( q \equiv (1 - \alpha) \mu / \eta, \epsilon \equiv 1 - \alpha, \) and \( \gamma \equiv \eta / (1 - \alpha)^2 \), we can rewrite this to

\[
\begin{align*}
\Delta w &= \gamma \epsilon [(1-\epsilon)q + \epsilon f(w, x)] \\
\Delta q &= \epsilon [f(w, x) - q].
\end{align*}
\]

We are interested in the evolution of the probability \( P(w, q, t) \) for the system to be in state \( (w, q) \) at time \( t \). With Poisson-distributed time intervals between succeeding learning steps, this probability \( P(w, q, t) \) obeys the master equation [11, 2]

\[
\frac{\partial P(w, q, t)}{\partial t} = \int dw' dq' [T(w, q \mid w', q') P(w', q', t) - T(w', q' \mid w, q) P(w, q, t)],
\]

with transition probability

\[
T(w, q \mid w', q') = \langle \delta(w - w' - \gamma \epsilon (1 - \epsilon) q + \epsilon f(w', x)) \delta(q - q' - \epsilon [f(w', x) - q']) \rangle_{\Omega},
\]

where \( \langle \cdot \rangle_{\Omega} \) denotes an average over the set \( \Omega \) of training patterns. Averages with respect to the probability density \( P(w, m, t) \) will be indicated by \( \langle \cdot \rangle_{\Xi} \) or, more explicitly, by \( \langle \cdot \rangle_{\Xi(w)} \). The master equation (4) is the starting point of the theoretical analysis presented in this paper. For notational convenience we treat the weight vector \( w \) as a one-dimensional variable. Generalisation to higher dimensions is straightforward and has no influence on the basic ideas presented in this paper.
1.3 Outline

In section 2 we will study the system (3) for finite $\gamma$ in the limit of very small $\epsilon$, i.e., for small learning parameters $\eta$ and momentum parameters $\alpha$ close to 1. In this case the time scales of the equations for the weight $w$ and the auxiliary variable $q$ are of the same order. We can immediately apply Van Kampen’s expansion [12] to the master equation (4) and obtain evolution equations for the average weight $w$ and the fluctuations around this average.

The situation $\epsilon$ finite and $\gamma$ small, which corresponds to finite momentum parameters $\alpha$ (not close to 1) and (again) small learning parameters $\eta$, will be considered in section 3. Now the evolution of the auxiliary variable $q$ takes place on a much faster time scale than the evolution of the weight $w$. Through projection of the master equation (4) on the “slow” space of the weight $w$, the fast variable $q$ can be eliminated, resulting again in (approximate) evolution equations for the weight $w$.

In section 4 we check our theoretical results with simulations of the nonlinear Oja learning rule [13]. The main results are summarised and discussed in section 5.

2 Equal time scales

2.1 Van Kampen’s expansion

In this section we will study the two-dimensional system (3) for small values of $\epsilon$ and finite values of $\gamma$, i.e., in the limits $\eta \rightarrow 0$ and $\alpha \rightarrow 1$ with a constant ratio $\gamma = \eta/(1 - \alpha)^2$. The master equation (4) can be approximated for small parameters $\epsilon$ using Van Kampen’s expansion. Basically (see [12, 5, 6] for a more detailed description of Van Kampen’s expansion), this expansion is based on the assumption that the stochastic process (3) can be viewed as a deterministic trajectory with (small) superimposed fluctuations of order $\sqrt{\epsilon}$. Starting from the Ansätze

\[ w = \phi + \sqrt{\epsilon} \xi \quad \text{and} \quad q = \psi + \sqrt{\epsilon} \chi, \]

Van Kampen’s expansion yields evolution equations for the deterministic variables $\phi$ and $\psi$, and for the average and (co)variance of the noise terms $\xi$ and $\chi$.

After rescaling time with $\gamma \epsilon$ (we define a new time $\tau \equiv \gamma \epsilon t$), we obtain the deterministic equations

\[ \begin{cases} \dot{\phi} = \psi \\ \gamma \dot{\psi} = f_1(\phi) - \psi, \end{cases} \tag{5} \]

with drift $f_1(\phi)$, the first moment of the learning rule $f(\phi, x)$. For later purposes we give the general definition of the $k$-th jump moment:

\[ f_k(\phi) \equiv \left\langle f^k(\phi, x) \right\rangle_\Omega. \tag{6} \]

The evolution of the averages of the noise terms follows

\[ \gamma \frac{d}{d\tau} \begin{pmatrix} \langle \xi \rangle \Xi \\ \langle \chi \rangle \Xi \end{pmatrix} = -A(\phi) \begin{pmatrix} \langle \xi \rangle \Xi \\ \langle \chi \rangle \Xi \end{pmatrix}, \]

with

\[ A(\phi) = \begin{pmatrix} 0 & -\gamma \\ -f'_1(\phi) & 1 \end{pmatrix}, \tag{7} \]

where the prime denotes differentiation of the function with respect to its argument. All learning networks are initialised at the same weight configuration, i.e., $w(0) = \phi(0)$ for all networks in
the ensemble \( \Xi \). This immediately implies \( \langle \xi \rangle_\Xi(t) = \langle \chi \rangle_\Xi(t) = 0 \) for all later times \( t \). From (5) we then derive that the average network state \( \langle w \rangle_\Xi = \phi \) obeys the second-order differential equation

\[
\gamma \ddot{\phi} + \dot{\phi} - f_1(\phi) = 0.
\]

The evolution of the covariance matrix

\[
\Sigma^2 \equiv \begin{pmatrix}
\langle \xi^2 \rangle_\Xi & \langle \xi \chi \rangle_\Xi \\
\langle \xi \chi \rangle_\Xi & \langle \chi^2 \rangle_\Xi
\end{pmatrix}
\]

is governed by

\[
\gamma \frac{d}{dt} \Sigma^2 = -A(\phi) \Sigma^2 - \Sigma^2 A(\phi) + D(\phi, \psi),
\]

with diffusion matrix

\[
D(\phi, \psi) \equiv \begin{pmatrix}
\gamma^2 \psi^2 & \gamma \psi [f_1(\phi) - \psi] \\
\gamma \psi [f_1(\phi) - \psi] & f_2(\phi) - 2 \psi f_1(\phi) + \psi^2
\end{pmatrix}.
\]

The \textit{a priori} Ansatz in Van Kampen’s expansion is that the noise terms \( \xi \) and \( \chi \) are of order 1. From (7) and (8), we see that this is valid for short times \( t \) and in regions of weight space where the real parts of the eigenvalues of the matrix \( A(\phi) \) are positive, i.e., where \( f_1'(\phi) < 0 \). The same conditions hold for the validity of Van Kampen’s expansion of the plain learning process (1) [5, 6].

### 2.2 Scaling properties

Let us take a closer look at the evolution equations for the average network state and the fluctuations around this average. With definitions

\[
\sigma_1 \equiv \frac{1}{\gamma} \langle \xi^2 \rangle_\Xi, \quad \sigma_2 \equiv \frac{1}{\gamma} \langle \xi \chi \rangle_\Xi, \quad \text{and} \quad \sigma_3 \equiv \langle \chi^2 \rangle_\Xi,
\]

the evolution equations (5) and (8) can be rewritten to

\[
\begin{align*}
\dot{\phi} - \psi &= 0, \\
f_1(\phi) - \psi &= \gamma \dot{\psi}, \\
\dot{\sigma}_1 - 2 \sigma_2 - \psi^2 &= 0, \\
f_1'(\phi) \sigma_1 - \sigma_2 + \sigma_3 + \psi [f_1(\phi) - \psi] &= \gamma \dot{\sigma}_2, \\
-2 \sigma_3 + f_2(\phi) - 2 \psi f_1(\phi) + \psi^2 &= \gamma \dot{\sigma}_3 - 2 \gamma f_1'(\phi) \sigma_2.
\end{align*}
\]

In this set of coupled differential equations, \( \gamma \) is the only remaining parameter. Suppose we know, through calculations or simulations, \( \phi(\tau) \) and \( \sigma_1(\tau) \) for a particular value of \( \gamma = \eta/(1 - \alpha)^2 \). Then for all combinations \( (\eta, \alpha) \) with this particular \( \gamma \), the average weight and fluctuations at time \( t \) follow from (recall our definitions of time \( \tau \) and variance \( \sigma_1 \))

\[
\langle w \rangle_\Xi(t) = \phi(\bar{\eta} t) \quad \text{and} \quad \langle w^2 - \langle w \rangle^2 \rangle_\Xi(t) = \bar{\eta} \sigma_1(\bar{\eta} t),
\]

with “rescaled learning parameter” \( \bar{\eta} \equiv \eta/(1 - \alpha) \). This rescaled learning parameter regulates the trade-off between speed and accuracy: a twice as large rescaled learning parameter leads to a twice as fast time scale, but also doubles the fluctuations in the weights. In section 4 we will describe simulations with the nonlinear Oja learning rule to check these scaling properties.

For small \( \gamma \) we can further simplify the set of equations (9). There are two different time scales: a slow time scale for the evolution of \( \phi \) and \( \sigma_1 \) and a fast time scale for the evolution of
\(\psi, \sigma_2, \sigma_3\). If we neglect all terms of order \(\gamma\), we can eliminate the fast variables \(\psi, \sigma_2\) and \(\sigma_3\) and obtain

\[
\begin{align*}
\dot{\phi} &= f_1(\phi) \\
\dot{\sigma}_1 &= 2f'_1(\phi)\sigma_1 + f_2(\phi).
\end{align*}
\] (11)

The same set of equations is obtained if Van Kampen’s expansion is applied to the plain learning rule (1) with rescaled learning parameter \(\tilde{\eta} = \eta/(1 - \alpha)\) (see e.g. \([5, 6]\)). Similar results have been reported in earlier studies on linear learning rules \([9, 10, 14]\). In the next section we will generalise these results to nonlinear learning rules for any finite value of the momentum parameter \(\alpha\), i.e., not close to 1. There we will go the other way around: first we will have to eliminate the fast variable \(q\) and only then can we apply Van Kampen’s expansion.

### 3 Different time scales

#### 3.1 Perturbation theory

In this section we will study the master equation (4) for small values of \(\gamma\) and finite values of \(\epsilon\), i.e., for small learning parameters \(\eta\) and momentum parameters \(\alpha\) not close to 1. In these limits, we cannot approximate the master equation by Van Kampen’s expansion as in section 2. However, as the results for \(\gamma \ll 1\) obtained in the previous section suggest, there are two different time scales in the master equation. In the long time limit, we can try to eliminate the fast variable \(q\) and then obtain (a series expansion of) an evolution equation for \(\dot{\tilde{P}}(w, t) \equiv \int dq\, P(w, q, t)\). Our approach is very loosely based on the “adiabatic elimination of fast variables” in the theory of stochastic processes \([15, 16]\).

Our starting point is the Kramers-Moyal expansion with respect to \(w\)

\[
\frac{\partial P(w, q, t)}{\partial t} = \left\{ n = 0 \right\} \left\{ \frac{(-\gamma\epsilon)^n}{n!} \frac{\partial^n}{\partial w^n} \int dq' \left\{ (\epsilon f(w, x) + (1 - \epsilon)q')^n \right. \right.
\]
\[
\times \left. \delta (q - (\epsilon f(w, x) + (1 - \epsilon)q')) \right\} \left( \int dq' P(w, q', t) \right) - P(w, q, t),
\] (12)

which is a completely equivalent representation of the master equation (4) (see e.g. \([12, 16]\)). From this Kramers-Moyal expansion we will derive evolution equations for the moments

\[
Q_k(w, t) \equiv \int dq\, q^k P(w, q, t), \quad k = 0, \ldots, \infty.
\]

Note that the moment vector \(\tilde{Q}(w, t)\) is just a different representation of the probability distribution \(P(w, q, t)\) and that \(Q_0(w, t) = \tilde{P}(w, t)\). Multiplying (12) by \(q^k\) and integrating over \(q\) yields (recall our definitions \(\epsilon = 1 - \alpha\) and \(\tilde{\eta} = \eta/(1 - \alpha)\)).

\[
\frac{\partial Q_k(w, t)}{\partial t} = \sum_{n=0}^{\infty} \frac{(-\tilde{\eta})^n}{n!} \frac{\partial^n}{\partial w^n} \sum_{l=0}^{n+k} \binom{n+k}{l} (1 - \alpha)^{n+k-l} f_{n+k-l}(w) \alpha^l Q_l(w, t) - Q_k(w, t),
\] (13)

with the jump moments \(f_k(w)\) defined in (6). We can write equation (13) as a formal evolution equation (for notational convenience we suppress the \(w\) and \(t\) dependence):

\[
\frac{\partial}{\partial t} \tilde{Q} = H \tilde{Q}
\] (14)

with

\[
H = \sum_{n=0}^{\infty} \tilde{\eta}^n H^{(n)},
\] (15)
in which the matrices $H^{(n)}$ are defined component wise by the operators

$$H^{(n)}_{ij} = \left( (-1)^n \frac{\partial^n}{\partial w^n} \sum_{i=0}^{n+i} \binom{n+i}{i} (1 - \alpha)^n \alpha^i \delta_{ij} \right) - \delta_{n0}\delta_{ij}, \quad i, j = 0, 1, \ldots, \infty. \quad (16)$$

The fact that the operator $H$ can be written as a series in the small parameter $\eta$ [equation (15)] gives us the possibility to treat the system (14) using perturbation theory.

Let us first consider the unperturbed ($\eta = 0$) system

$$\frac{\partial}{\partial t} \tilde{Q} = H^{(0)} \tilde{Q}. \quad (17)$$

From the triangular form of $H^{(0)}$, we immediately find its degenerate eigenvalues

$$\lambda^{(0)}_\kappa = -(1 - \alpha^\kappa), \quad \kappa = 0, 1, \ldots, \infty.$$  

We define $V^{(0)}_\kappa$ as the subspaces of eigenvectors with eigenvalue $\lambda^{(0)}_\kappa$, and $P^{(0)}_\kappa$ as the orthogonal projectors (i.e., $P^{(0)}_\kappa P^{(0)}_\mu = \delta_{\kappa\mu} P^{(0)}_\mu$) on the subspaces $V^{(0)}_\kappa$. These projectors commute with $H^{(0)}$, i.e.,

$$P^{(0)}_\kappa H^{(0)} = H^{(0)} P^{(0)}_\kappa = -(1 - \alpha^\kappa) P^{(0)}_\kappa. \quad (18)$$

The projection $[P^{(0)}_\kappa \tilde{Q}](w, t)$ is called a “mode.” From (18) it follows that the evolution of a mode is governed by

$$\frac{\partial}{\partial t} [P^{(0)}_\kappa \tilde{Q}] = H^{(0)} [P^{(0)}_\kappa \tilde{Q}] = -(1 - \alpha^\kappa) [P^{(0)}_\kappa \tilde{Q}].$$

Since the modes are independent, the solution of the unperturbed system (17) is the sum of the solution of the modes:

$$\tilde{Q}(w, t) = \sum_{\kappa=0}^{\infty} e^{-(1 - \alpha^\kappa)t} [P^{(0)}_\kappa \tilde{Q}](w, 0).$$

The modes with $\kappa \neq 0$ will rapidly relax to equilibrium. We call these modes the fast modes. For large $t$, only the slow mode, i.e., the one with $\kappa = 0$, will remain. We write $P^{(0)}$ as the projector on the slow mode, i.e., $P^{(0)} \equiv P^{(0)}_0$. Consequently, the projector on the fast modes is $1 - P^{(0)}$. So, for large $t$ the fast modes will be equilibrated,

$$[1 - P^{(0)}] \tilde{Q} = 0, \quad (19)$$

and only the dynamics on the slow mode remains:

$$\frac{\partial}{\partial t} P^{(0)} \tilde{Q} = H^{(0)} P^{(0)} \tilde{Q}. \quad (20)$$

It is illustrative to see how we can arrive at an evolution equation for $\hat{P}(w, t)$ from equation (20) using properties of the projector $P^{(0)}(w)$ and the operator $H(w)$. In the appendix it is shown that the projector $P^{(0)}(w)$ has components

$$P^{(0)}_{ij}(w) = v_i(w) k_{0j},$$

where the vector $\vec{v}(w)$ obeys $H^{(0)}(w)\vec{v}(w) = 0$ and $v_0(w) = 1$. Using the constraint (19), we can express all components $Q_k(w, t)$ in terms of the zeroth component $Q_0(w, t)$: $Q_k(w, t) = v_k(w) Q_0(w, t)$. This corresponds to elimination of the fast variable $q$ and can be compared with the elimination of the variables $\psi, \sigma_2,$ and $\sigma_3$ in equation (9). Equation (20) now reduces to an equation for $Q_0(w, t) = \hat{P}(w, t)$ only:

$$\frac{\partial}{\partial t} \hat{P}(w, t) = 0.$$
This equation makes the rather trivial statement that in the unperturbed system ($\tilde{\eta} = 0$) no learning takes place.

For the perturbed system (14) we follow the same line of reasoning as for the unperturbed system (17). The starting point of perturbation theory is the assumption that the eigenvalues and eigenvectors of the perturbed system can be written as an expansion in the perturbation parameter $\tilde{\eta}$. We define $V_\kappa$ as the subspaces spanned by the eigenvectors corresponding to eigenvalues of which the unperturbed value is $\lambda_\kappa^{(0)}$, and $P_\kappa$ as the orthogonal projectors on these subspaces $V_\kappa$. As in the unperturbed case, we decompose the perturbed system into modes.

The eigenvalues with $\kappa = 0$ are of order $\tilde{\eta}$, whereas the eigenvalues with $\kappa \neq 0$ are equal to $-(1 - \alpha^\kappa)$ plus terms of order $\tilde{\eta}$. So, if $\tilde{\eta} \ll 1 - \alpha$, the eigenvalues with $\kappa = 0$ are much smaller in absolute value than the eigenvalues with $\kappa \neq 0$, and we can still distinguish the slow mode from the fast modes. Again, we use the abbreviation $P = P_\kappa = 0$ for the projector on the slow mode. For large $t$, the fast modes will be equilibrated, i.e.,

$$[1 - P] \tilde{\eta}^0 = 0,$$

and only the dynamics on the slow mode remains:

$$\frac{\partial}{\partial t} P \tilde{Q} = Hp \tilde{Q}.$$  

Due to the constraint (21), all the components $Q_0$ of $\tilde{Q}$ are determined once $Q_0$ is known. In other words, we can use the constraint (21) to derive a dynamical equation for $Q_0(w, t) = \tilde{P}(w, t)$ from equation (22). In this way, the fast variable $q$ is eliminated from the master equation.

Since the operator $H$ is only known in the form of a series expansion, the best we can achieve is a series expansion of the evolution equation for $Q_0$ in powers of $\tilde{\eta}$. In order to obtain this series expansion we only have to consider one of the components of (22). The zeroth component

$$\frac{\partial}{\partial t} (P \tilde{Q})_0 = (Hp \tilde{Q})_0$$

is the most obvious choice. The unknown quantities in this equation are both $\tilde{Q}$ and the projector $P$. Writing $P$ as a series

$$P = \sum_{n=0}^{\infty} \tilde{\eta}^n P^{(n)}$$

we can subtract the desired components of $P^{(n)}$ from the properties $P^2 = P$ and $Hp = PH$.

Using (21) and (23) we can then express the components $\tilde{Q}_k$ with $k \neq 0$ in terms of the zeroth component $Q_0$ and derive an evolution equation for $Q_0 = \tilde{P}$ to arbitrary order in $\tilde{\eta}$. In the appendix it is shown that this expansion yields

$$\frac{\partial}{\partial t} \tilde{P}(w, t) = -\tilde{\eta} \frac{\partial}{\partial w} f_1(w) \tilde{P}(w, t) + \frac{\tilde{\eta}^2}{2} \frac{\partial^2}{\partial w^2} f_2(w) \tilde{P}(w, t)$$

$$+ \frac{\tilde{\eta}^2 \alpha}{1 - \alpha} \left[ \frac{\partial^2}{\partial w^2} f_1(w) \tilde{P}(w, t) - \frac{\partial}{\partial w} f_1(w) \frac{\partial}{\partial w} f_1(w) \tilde{P}(w, t) \right] + O(\tilde{\eta}^3).$$

3.2 Van Kampen’s expansion

To study (24) in the limit $\tilde{\eta} \rightarrow 0$, we apply Van Kampen’s expansion [12]. We start with the Ansatz

$$w = \phi(\tau) + \sqrt{\tilde{\eta}} \zeta,$$

where $\tau = \tilde{\eta} t$ and $\phi(\tau)$ is a function to be determined (compare with section 2.1). Note that the constraint (21) and thus the evolution equation (24) are valid for times $t = O(1/\tilde{\eta})$, i.e., for $\tau = O(1)$. The function $\Pi(\zeta, \tau)$ is the probability $\tilde{P}$ in terms of the new variable $\zeta$:

$$\Pi(\zeta, \tau) \equiv \tilde{P}(\phi(\tau) + \sqrt{\tilde{\eta}} \zeta, \tau/\tilde{\eta}).$$
From Van Kampen’s expansion it immediately follows that the deterministic part $\phi(\tau)$ has to satisfy the equation

$$\frac{d\phi(\tau)}{d\tau} = f_1(\phi(\tau))$$

(26)

and the evolution of $\Pi(\zeta, \tau)$ is governed by the Fokker-Planck equation

$$\frac{\partial \Pi(\zeta, \tau)}{\partial \tau} = -f'_1(\phi(\tau)) \frac{\partial}{\partial \zeta} \Pi(\zeta, \tau) + \frac{1}{2} f_2(\phi(\tau)) \frac{\partial^2}{\partial \zeta^2} \Pi(\zeta, \tau).$$

(27)

The solution of the Fokker-Planck equation (27) is a Gaussian, so it suffices to determine the first and the second moments of $\zeta$:

$$\frac{d\langle \zeta \rangle_{\Xi}}{d\tau} = f'_1(\phi(\tau)) \langle \zeta \rangle_{\Xi}$$

$$\frac{d\langle \zeta^2 \rangle_{\Xi}}{d\tau} = 2 f'_1(\phi(\tau)) \langle \zeta^2 \rangle_{\Xi} + f_2(\phi(\tau)).$$

(28)

From these equations, we see that the fluctuations $\zeta$ are bounded if $f'_1(\phi(\tau)) < 0$. If $f$ satisfies this condition, the Ansatz (25) is a posteriori justified. On the other hand, in case of non-negative $f'_1(\phi(\tau))$, the fluctuations grow in time and the expansion need not to be valid. As in section 2.1, this condition on $f$ does not depend on $\alpha$. Note further that $\langle \zeta^2 \rangle_{\Xi} = \sigma^2$, the variance defined in section 2.2: the equations (26) and (27) with $\tilde{\eta} = \eta$ substituted. In the first place, this result is a verification of our analysis, since learning without momentum is learning with $\alpha = 0$ and $\tilde{\eta} = \eta$. In the second place, the result shows that for learning parameters $\eta \ll (1 - \alpha)^2$, from the point of view of the Fokker-Planck approximation, learning with momentum term is equivalent to learning without momentum term with rescaled learning parameter $\tilde{\eta}$.

4 Simulations

To illustrate the analytical results of the previous sections, we simulate the process of on-line learning with a momentum term for the nonlinear learning rule of Oja [13] in two dimensions

$$\Delta w(n) = \eta (x^T w(n)) [x - (x^T w(n)) w(n)] + \alpha \Delta w(n - 1).$$

Oja’s rule searches for the principal component of the input correlation matrix $\langle x x^T \rangle_{\Omega}$. Inputs $x$ are drawn at random from a rectangle centered at the origin, with sides of length 2 and 1 along the $x_1$ and $x_2$-axis, respectively. Simulations are performed with an ensemble of 100 000 independently learning networks. The networks in the ensemble are asynchronously updated. This means that at each step only one, randomly chosen network in the ensemble is updated. Hence, for a single network in the ensemble, the time intervals between updates are binomially distributed. For a large ensemble this distribution approaches a Poisson distribution [17]. The time scale $t$ is such that there is on the average one learning step per unit of time for each network in the ensemble. All networks are initialised at the weight configuration $w(0) = (0.3, 0.3)^T$. Since the principal component of the input correlation matrix lies along the longest side of the rectangle, the weights $w_1$ and $w_2$ tend to 1 and 0, respectively. In figures 1 and 2 we plot the evolutions of the average weights $\langle w \rangle_{\Xi(t)}$ and of the trace of the covariance matrix

$$\sigma^2(t) = \sum_{i=1}^{2} \langle (w_i - \langle w_i \rangle_{\Xi(t)})^2 \rangle_{\Xi(t)}$$

for various values of $\alpha$ and $\eta$. 

In section 2 we derived that, for small learning parameters $\eta$ and momentum parameters $\alpha$ close to 1, all combinations $(\eta, \alpha)$ with the same value of $\gamma = \eta/(1 - \alpha)^2$ give rise to similar behaviour. We verify this scaling property in figure 1. In each graph we keep $\gamma$ constant ($\gamma = 0.1$, 1, and 5 for figure 1(a), (b), and (c), respectively) and present curves for different values of $\alpha$ ($\alpha = 0.9$, 0.8, and 0.6 for solid, dashed, and dash-dotted lines, respectively). Time and variance are rescaled with $\tilde{\eta}$, i.e., we plot

$$\phi(\tau) = <w>_{\Xi(\tau/\tilde{\eta})} \quad \text{and} \quad \sigma_1(\tau) = \sigma^2(\tau/\tilde{\eta})/\tilde{\eta}$$

as functions of the rescaled time $\tau$ [compare with equation (10)]. Curves with equal $\gamma$ are almost overlapping, except for the quite extreme values $\alpha = 0.6$ and $\eta = 0.8$ [dash-dotted lines in figure 1(c)]: the simulation results are in perfect agreement with the scaling properties derived in section 2.
In section 3 we considered the case $\eta \ll (1 - \alpha)^2$, i.e., $\gamma \ll 1$, and showed that combinations $(\eta, \alpha)$ are equivalent to combinations $(\tilde{\eta}, 0)$. This claim is verified in figure 2. In figures 2(a) and (b) the rescaled learning parameter $\tilde{\eta}$ is kept constant ($\tilde{\eta} = 0.01$ and 0.1, respectively), thus there is no need to rescale time and variance. Each graph shows curves with different values of $\alpha$ ($\alpha = 0, 0.5,$ and 0.9 for solid, dashed, and dash-dotted lines, respectively), i.e., different values of $\gamma$. Curves inside each graph are almost overlapping, even when the values of $\gamma$ differ by a factor 10 in magnitude [solid and dash-dotted line in figure 2(a)]. The exception is the dashed-dotted line in figure 2(b) where $\gamma = 1$ is not small enough for our analysis to be valid. We conclude that these simulation results agree very well with the analytical results of section 3.

5 Discussion

In this paper we studied nonlinear on-line learning rules with momentum term for small learning parameters $\eta$. We considered two cases: momentum parameters $\alpha$ close to 1, and finite momentum parameters $\alpha$. In the first case we took the limits $\eta \rightarrow 0$ and $\alpha \rightarrow 1$, keeping $\gamma = \eta/(1 - \alpha)^2$ constant. Using Van Kampen’s expansion we arrived at evolution equations for the average weight vector and the fluctuations around this average. These evolution equations depend (after proper rescaling of time and size of fluctuations) only on this parameter $\gamma$. In the second case we kept $\alpha$ constant and again took the limit $\eta \rightarrow 0$. We arrived at the conclusion that learning with learning parameter $\eta$ and momentum parameter $\alpha$ is equivalent to learning with rescaled learning parameter $\tilde{\eta} = \eta/(1 - \alpha)$. Note that exactly the same conclusion follows from the first case in the limit $\gamma \rightarrow 0$. Thus, the analysis of the second case shows that the set of equations (9)
resulting from the first case can be used not only for momentum parameters $\alpha$ close to 1, but also for more general values of $\alpha$.

We tried to answer the question whether incorporation of the momentum parameter really improves the performance of general on-line learning rules. For learning parameters $\eta \ll (1 - \alpha)^2$, we found that the effect of the momentum term is nothing but a rescaling of the learning parameter. For practical applications, this result is quite disappointing, but in agreement with the notion that the momentum term is hardly ever used in combination with on-line learning rules. The important exception is backpropagation. Indeed, it can be argued that incorporation of the momentum term is helpful for batch-mode backpropagation (see e.g. [9, 10, 18]), but neither these arguments, nor the results presented in this paper can explain the popularity of on-line backpropagation with momentum updating. There might be several reasons for this. Our analysis holds only for small learning parameters $\eta$ and in regions of weight space in the vicinity of local minima, so, it could be that our analysis is too restricted. Another option is that the momentum term only helps if the momentum parameter $\alpha$ and learning parameter $\eta$ are chosen such that $\gamma = \eta/(1 - \alpha)^2 = O(1)$. Analysis of the evolution equations for linear learning rules, that can be solved for any finite value of $\alpha$ and $\eta$ and are valid in the whole weight space, do not show any significant improvement of on-line learning with momentum term if compared to learning without momentum term (unpublished results). For stronger evidence we will have to come up with a more general analysis of nonlinear learning with momentum updating and/or to work towards a better understanding of the nonlinear set of equations (9). At this point, we tend to the conclusion that the popularity of the momentum term in combination with on-line backpropagation cannot be explained in mathematical terms, but perhaps better in psychological terms: on-line backpropagators are afraid to choose a large learning parameter themselves.

Acknowledgments

This work was supported by the Dutch Foundation for Neural Networks (WW and AK) and by a grant (P41RR05969) from the National Institutes of Health (TH). We thank Karin Krommenhoek and Bert Kappen for stimulating discussions.

Appendix

In this appendix, we will show how to derive the evolution equation (24) for $\hat{P}(w,t)$, starting from the evolution equation (14) for the moment vector $\hat{Q}(w,t)$. Since we are interested in $\hat{P}(w,t) = Q_0(w,t)$, we consider the zeroth component of

$$\frac{\partial}{\partial t} (\mathcal{P}\hat{Q})_0 = (H\mathcal{P}\hat{Q})_0$$

under the constrains

$$(1 - \mathcal{P})\hat{Q} = 0.$$  \hfill (A.2)

The operator $H$ is defined in (15) and (16) and the projector $\mathcal{P}$ is written as a series expansion in (23). The unknown factors in (A.1) and (A.2) are not only the elements of the vector $\hat{Q}$, but also the components of the corrections $\mathcal{P}^{(n)}$ of the projector. The latter ones will be deduced from the relations $\mathcal{P}^2 = \mathcal{P}$ and $H\mathcal{P} = \mathcal{P}H$.

We expand both sides of (A.1) to the two lowest orders in $\hat{\eta}$:

$$\frac{\partial}{\partial t} \sum_{j=0}^{\infty} \left\{ \hat{p}^{(0)}_{0,j} + \hat{\eta}\hat{p}^{(1)}_{0,j} + \mathcal{O}(\hat{\eta}^2) \right\} Q_j$$

$$= \sum_{j=0}^{\infty} \left\{ \hat{\eta}(H^{(1)}\mathcal{P}^{(0)})_{0,j} + \hat{\eta}^2 \{ (H^{(2)}\mathcal{P}^{(0)})_{0,j} + (H^{(1)}\mathcal{P}^{(1)})_{0,j} \} + \mathcal{O}(\hat{\eta}^3) \right\} Q_j, \quad \text{(A.3)}$$
where we used that \( H_{0j}^{(0)} = 0 \) \( \forall j \). Note that there is a global scale factor \( \tilde{\eta} \) on the right hand side. This global scale factor will later be incorporated in a rescaled time. From (16) it follows that only the first \((n + 1)\)-th components of the zeroth row of \( H^{(n)} \) can be nonzero. Therefore we only need to calculate the first two rows \( P_{0j}^{(1)} \) and \( P_{1j}^{(1)} \), and the first three rows \( P_{0j}^{(0)} \), \( P_{1j}^{(0)} \) and \( P_{2j}^{(0)} \).

First, we calculate the components of the unperturbed projector \( P^{(0)} \). Using \( h_{P_{ij}^{(0)}}^2 = P_{ij}^{(0)} \) and \( P_{ij}^{(0)} H^{(0)} = H^{(0)} P_{ij}^{(0)} = 0 \) we find that \( P^{(0)} \) has components

\[
P_{ij}^{(0)} = \delta_{ij0} \]

where \( \tilde{v} \) is the vector satisfying \( H^{(0)} \tilde{v} = 0 \), with the function-valued components

\[
v_0 = 1, \\
v_1 = f_1, \\
v_2 = (1 - \alpha) f_2 + 2 \alpha f_1^2 / (1 + \alpha), \\
v_3 = \ldots .
\]

Now we consider the first correction \( P^{(1)} \). From \( P^2 = P \), the first correction \( P^{(1)} \) should satisfy

\[
P^{(1)} = P^{(0)} P^{(1)} + P^{(1)} P^{(0)}.
\]

For the components of \( P^{(1)} \) this implies

\[
P_{ij}^{(1)} = \sum_{k=0}^{\infty} P^{(1)}_{ik} v_k \delta_{jk0} + v_i P_{0j}^{(0)} ,
\]

which, after some rewriting, yields

\[
\sum_{k=0}^{\infty} v_i P^{(1)}_{0k} v_k = 0 \quad \text{and} \quad P_{ij}^{(1)} = v_i P_{0j}^{(1)} \quad \text{for} \quad j \neq 0 .
\]

(A.4)

Since \( P \) commutes with \( H \), \( P^{(1)} \) should also obey

\[
H^{(1)} P^{(0)} + H^{(0)} P^{(1)} = P^{(0)} H^{(1)} + P^{(1)} H^{(0)} .
\]

(A.5)

Using the explicit forms of \( H^{(0)} \), \( H^{(1)} \) and \( P^{(0)} \), we deduce from (A.4) and (A.5) that

\[
P_{00}^{(1)} = \alpha \frac{\partial}{1 - \alpha \partial w} f_1 , \\
P_{01}^{(1)} = -\alpha \frac{\partial}{1 - \alpha \partial w} , \\
P_{0k}^{(1)} = 0 \quad \text{for} \quad k \geq 2 , \\
P_{10}^{(1)} = \frac{1 + \alpha}{1 - \alpha} \frac{\partial}{\partial w} f_1 - \frac{1}{(1 - \alpha^2) \partial w} \{ (1 - \alpha) f_2 + 2 \alpha f_1^2 \} , \\
P_{11}^{(1)} = -\frac{\alpha}{1 - \alpha} \frac{\partial}{\partial w} f_1 , \\
P_{1k}^{(1)} = 0 \quad \text{for} \quad k \geq 2 .
\]

(A.6)

The constraint (A.2) gives the relation between the zeroth and the first component of \( \tilde{Q} \):

\[
Q_1 = f_1 Q_0 + \mathcal{O}(\tilde{\eta}) .
\]

(A.7)
Substitution of the expansions (16), (A.6) and (A.7) for the operator $H$, the projector $P$ and the moment vector $\tilde{Q}$, respectively, into the evolution equation (A.1) finally leads to

$$\frac{\partial}{\partial t} Q_0 = \left\{ -\tilde{\eta} \frac{\partial}{\partial w} f_1 + \frac{\tilde{\eta}^2}{2} \frac{\partial^2}{\partial w^2} f_2 + \frac{\tilde{\eta}^2 \alpha}{1 - \alpha} \left[ \frac{\partial^2}{\partial w^2} f_2 - \frac{\partial}{\partial w} f_1 \frac{\partial}{\partial w} f_1 \right] \right\} Q_0 + O(\tilde{\eta}^3),$$

which, after substitution of $Q_0(w,t) = \tilde{P}(w,t)$, is the desired result (24).

References


