Neural networks learning in a changing environment

Tom Heskes & Bert Kappen

Department of Medical Physics and Biophysics
University of Nijmegen
Geert Grooteplein Noord 21
6525 EZ Nijmegen, The Netherlands
e-mail: tom@mbfys.kun.nl

Abstract

We study the learning dynamics of neural networks from a general point of view. A learning algorithm that enables a neural network to adapt to a changing environment, must have a non-vanishing learning parameter. This constant adaptability, however, goes at the cost of the accuracy, i.e. the size of the fluctuations in the plasticities, such as synapses and thresholds. In some cases an optimal learning parameter can be calculated.

1 Introduction

In neural network models, learning plays an essential role. Learning is the mechanism by which a network adapts itself to its environment. So far the learning process in artificial neural networks has been considered almost exclusively for the case when the network is given examples from a *fixed* environment. The environment can be defined as a set of examples or stimuli, and learning is usually modelled as the process of randomly drawing examples from the environment and presenting them to the neural network. Thus learning becomes a stochastic process. An ensemble theory, describing the evolution of an ensemble of learning neural networks, is necessary. In order the prevent fluctuations in the asymptotic state, the learning parameter, which controls the amount of learning, should go to zero for large times.

Such algorithms, for which the learning parameter vanishes asymptotically, are clearly not the ones that are used in natural neural networks. Natural adaptive systems always learn. This constant tendency to learn accounts for the adaptability of biological neural systems to a *changing* environment. In order to implement such behavior in artificial neural networks, the learning parameter should take a constant nonzero value. Therefore, we propose to study the learning dynamics of a large class of neural networks for constant learning parameters.

0-7803-0164-1/91/0000-0823$01.00©1991 IEEE

2 Learning in a changing environment

The state of the neural network is denoted by the N-dimensional vector $\mathbf{w} = (w_1, \ldots, w_N)^T$, including all synapses and thresholds. The environment Ω of the network is assumed to be a set of n-dimensional stimuli $\vec{x}$ with corresponding probability distribution $\rho(\vec{x})$. We consider the following learning mechanism. At distinct points in time a stimulus $\vec{x}$ is drawn at random from Ω and is presented to the network. The network changes its weight vector $\mathbf{w}$ to $\mathbf{w} + \Delta\mathbf{w}$, obeying:

$$\Delta\mathbf{w} \;=\; \eta\, \mathbf{f}(\mathbf{w}, \vec{x}), \tag{1}$$

where $\mathbf{f}(\mathbf{w}, \vec{x})$ is called the stochastic force and η is the learning parameter. Eq. (1) simply states that the new network state $\mathbf{w}'$ after the learning step is a function of the state $\mathbf{w}$ before this learning step and the randomly drawn input vector $\vec{x}$. This applies to most of the learning rules in neural network theory, such as backpropagation [9], self-organization [8] [3] [7] and spin-glass learning rules [5] [2].

Learning, modelled as such, is a random walk in discrete iteration steps with transition probability

$$T(\mathbf{w}'|\mathbf{w}) \;=\; \int d\vec{x}\, \rho(\vec{x}) \delta\left(\mathbf{w}' - \mathbf{w} - \eta \mathbf{f}(\mathbf{w}, \vec{x})\right) \equiv \left\langle \delta\left(\mathbf{w}' - \mathbf{w} - \eta \mathbf{f}(\mathbf{w}, \vec{x})\right)\right\rangle_{\Omega}.$$

We introduce continuous time by drawing the time intervals Δt between two learning steps at random according to (the reason for this particular choice is explained in [1])

$$\varrho(\Delta t) \;=\; \frac{1}{\tau} \exp\left[-\frac{\Delta t}{\tau}\right].$$

The probability $P(\mathbf{w}, t)$ that the network is in state $\mathbf{w}$ at *time* t can be defined. $P(\mathbf{w}, t)$ obeys the continuous time master equation:

$$\tau \frac{dP(\mathbf{w}', t)}{dt} \;=\; \int d\mathbf{w}\, \left[T(\mathbf{w}'|\mathbf{w}) P(\mathbf{w}, t) \;-\; T(\mathbf{w}|\mathbf{w}') P(\mathbf{w}', t)\right] \tag{2}$$

In a gradually changing environment, such that that environmental changes on a time scale $\tau = \langle \Delta t \rangle$ are negligible, Eq. (2) still holds with $T(\mathbf{w}'|\mathbf{w})$ a smoothly varying function of time. The distribution of states $\mathbf{w}$ at time t is denoted by Ξ_t. The expression $\langle \Phi(\mathbf{w}) \rangle_{\Xi_t} \equiv \int d\mathbf{w}\, P(\mathbf{w}, t) \Phi(\mathbf{w})$ can be viewed as the average of the function $\Phi(\mathbf{w})$ over an ensemble of independently operating neural networks at time t.

Let us first give a short outline of the characteristics of learning processes in a *fixed* environment Ω. It can be shown [6] that, under certain conditions including a slowly vanishing of the learning parameter, the learning process converges to a stationary solution $P(\mathbf{w}, \infty) = \delta^N(\mathbf{w} - \mathbf{w}^*)$, where the points $\mathbf{w}^*$ are fixed points of the differential equation:

$$\frac{d\mathbf{w}(t)}{dt} \;=\; \langle \mathbf{f}(\mathbf{w}(t), \vec{x}) \rangle_{\Omega}.$$

If we keep the learning parameter at a small constant value, we can show [4] that the stationary probability distribution is peaked in the neighborhood of the fixed points $\mathbf{w}^*$. We can calculate the asymptotic value of the standard deviation matrix up to first order in η [4]:

$$\Sigma^2_{\infty} \equiv \left\langle \left(\mathbf{w} - \langle \mathbf{w} \rangle_{\Xi_\infty}\right) \left(\mathbf{w} - \langle \mathbf{w} \rangle_{\Xi_\infty}\right)^T \right\rangle_{\Xi_\infty} = \eta \int_0^{\infty} dy\, \mathrm{e}^{-Gy} D\, \mathrm{e}^{-G^T y}, \tag{3}$$

with the positive definite diffusion tensor $D_{ij} \equiv \langle f_i(\mathbf{w}^*, \vec{x}) f_j(\mathbf{w}^*, \vec{x}) \rangle_\Omega$, containing the fluctuations in the learning rule, and the positive definite matrix $G_{ij} \equiv -\partial_{w_j} \langle f_i(\mathbf{w}^*, \vec{x}) \rangle_\Omega$. If the matrix G is symmetric, we find:

$$\mathrm{Tr}\left[\Sigma_\infty^2\right] = \frac{\eta}{2}\mathrm{Tr}\left[G^{-1}D\right].$$

We conclude that for small learning parameters the remaining fluctuations are proportional to η, proportional to the fluctuations in the learning rule and inversely proportional to the curvature of the (local) energy surface at $\mathbf{w}^*$ of which G is the matrix of second derivatives.

In a *changing* environment the fixed points $\mathbf{w}^*(t)$ are, in general, a function of time. As an indication for the performance of an ensemble of neural networks in a changing environment, we define the error:

$$\begin{aligned} \mathcal{E} &= \lim_{T\to\infty} \frac{1}{T} \int_0^T dt \left\langle \|\mathbf{w} - \mathbf{w}^*(t)\|^2 \right\rangle_{\Xi_t} \\ &= \lim_{T\to\infty} \frac{1}{T} \int_0^T dt \left\{ \left\| \langle \mathbf{w} \rangle_{\Xi_t} - \mathbf{w}^*(t) \right\|^2 + \mathrm{Tr}\left[\Sigma_t^2\right] \right\}. \end{aligned} \tag{4}$$

There is an important conflict between the bias $\langle \mathbf{w} \rangle_{\Xi_t} - \mathbf{w}^*(t)$ and the standard deviation Σ_t^2 in Eq. (4). The bias stands for the *adaptibility* of the neural network. It can be shown [4] that, in a first approximation, it is proportional to the environmental change through $\dot{\mathbf{w}}^*$ and inversely proportional to the learning parameter. The leading term in the expansion for the fluctuations, representing the *accuracy* of a neural network, is, for slow environmental changes, proportional to the learning parameter as in Eq.(3). So, typically:

$$\mathcal{E} \approx \alpha \left[\frac{\dot{w}^*}{\eta}\right]^2 + \beta\eta,$$

where α and β are positive constants, independent of η and $\dot{w}^*$. Minimization leads to an optimal learning parameter

$$\eta_{optimal} \propto \left(\dot{w}^*\right)^{2/3}.$$

3 Two examples

Our first example is a Grossberg learning unit [3]

$$\Delta w = \eta\,(x - w),$$

trying to find the centre of mass of the input distribution: $w^*(t) = \langle x \rangle_{\Omega_t}$. The input distribution, the dotted box in Fig. 1, is moving with constant velocity $\dot{w}^*$ and standard deviation χ. The asymptotic macroscopic quantities $\langle w \rangle_{\Xi_t}$ and Σ_t^2 can be calculated exactly [4]:

$$\begin{aligned} \langle w \rangle_{\Xi_t} &= \langle x \rangle_{\Omega_t} - \frac{\tau\dot{w}^*}{\eta} = w^*(t - \tau/\eta) \\ \Sigma_t^2 &= \Sigma^2 = \frac{1}{\eta(2-\eta)} \left[(\eta\chi)^2 + (\tau\dot{w}^*)^2\right]. \end{aligned} \tag{5}$$

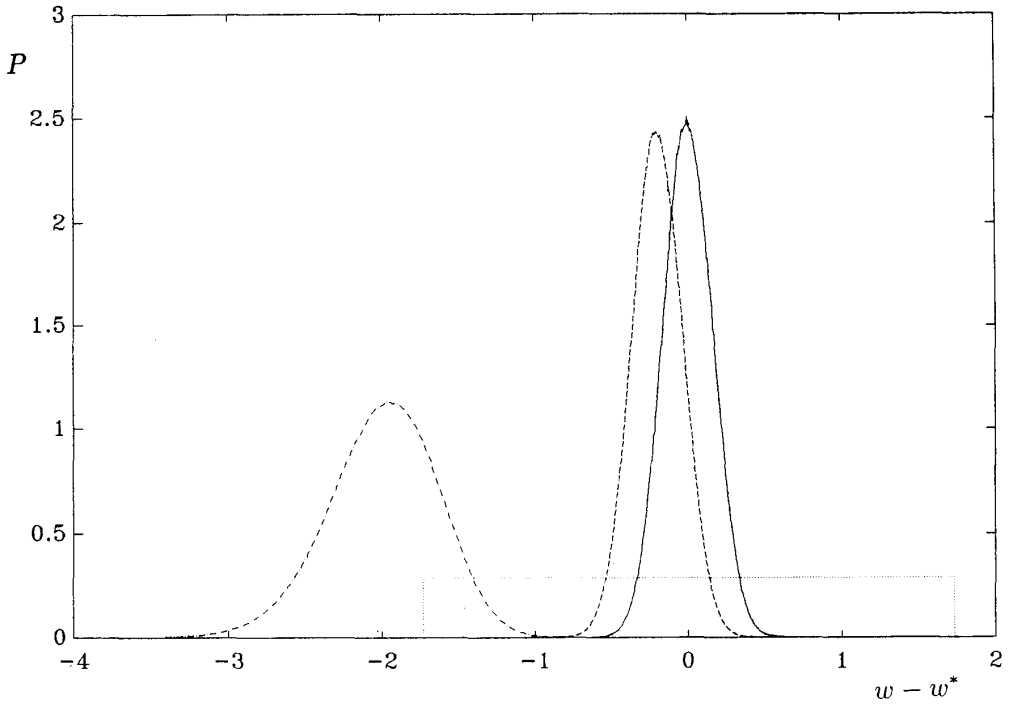

Figure 1: Simulated probability distribution for time dependent Grossberg learning. Learning parameter $\eta = .05$, standard deviation input $\chi = 1.0$, 5000 neural networks. Zero velocity (straight line). Small velocity: $\dot{w}^* = .01/\tau$ (dashed line). Relatively large velocity: $\dot{w}^* = .1/\tau$ (dashed-dotted line). The input distribution is plotted for reference (dotted line).

The probability distribution with respect to w^* is plotted for constant learning parameter and 3 different velocities in Fig. 1. The larger the velocity, the more the neural network lags behind and the broader the probability distribution. From Eq. (5) it is straightforward to compute the error and the optimal learning parameter.

The nonlinear Oja learning rule [7]

$$\Delta \mathbf{w} = \eta \, (\mathbf{w}^T \mathbf{x}) \, [\mathbf{x} - (\mathbf{w}^T \mathbf{x}) \mathbf{w}]$$

searches for the principal component of the covariance matrix $C(t) \equiv \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\Omega_t}$. The unit is given two dimensional examples from a rectangle as in Fig 2. The rectangle is rotating with angular velocity ω: $\mathbf{w}^*(t) = (\cos\omega t \, , \, \sin\omega t)^T$. The error yields approximately [4]:

$$\mathcal{E} = \frac{1}{\eta^2} \left[\frac{\omega\tau}{\Lambda_1 - \Lambda_2} \right]^2 + \frac{\eta}{2} \frac{\Lambda_1 \Lambda_2}{\Lambda_1 - \Lambda_2},$$

with Λ_1 and Λ_2 the eigenvalues of the covariance matrix ($\Lambda_1 > \Lambda_2$). In Fig. 3 this error is plotted as a function of the learning parameter η, both calculated and simulated. The difference between the computed and simulated values is due to the neglectance of higher order terms in η and ω. For $0.02 \leq \eta \leq 0.11$ the deviation is within 10%. A minimal error is, in this particular case, found for $\eta \approx 0.04$.

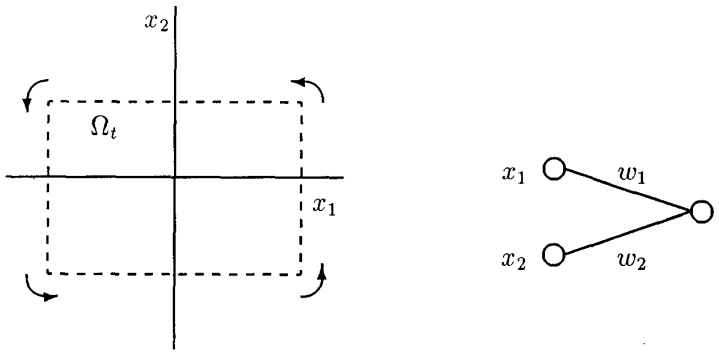

Figure 2: Oja learning. A unit is taught with 2-dimensional examples from a rectangle which is rotating around the origin. The principal component of the covariance matrix lies parallel to the longest side of the rectangle.

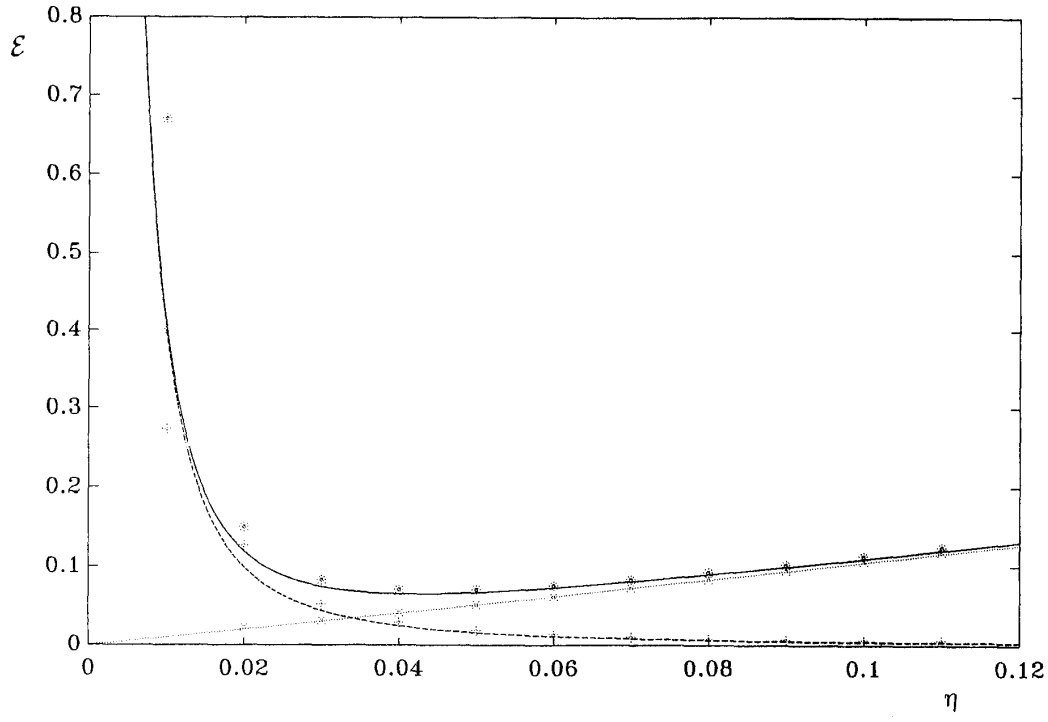

Figure 3: Squared bias, variance and error for time dependent Oja learning as a function of the learning parameter. Eigenvalues of the covariance matrix: $\Lambda_1 = 2$ and $\Lambda_2 = 1$. Angular velocity: $\omega = 2\pi/1000\tau$. Simulations were done with 5000 neural networks. Squared bias (computed: dashed line; simulated: '+'), variance (computed: dotted line; simulated: 'x') and error (computed: solid line; simulated: '$*$')

4 Conclusions

The introduction of Poisson distributed time steps facilitates a continuous time description of learning processes with non-vanishing learning parameters. We used this description to study the performances of neural networks operating in a changing environment. In a changing environment there is a trade-off between adaptibility and accuracy. The choice of the learning parameter for these networks deserves attention. Given a well-defined error, an optimal learning parameter can be estimated in some cases.

Acknowledgement

This work is partly supported by the Dutch Foundation for Neural Networks.

References

[1] D. Bedeaux, K. Lakatos-Lindenberg, and K. Shuler. On the relation between master equations and random walks and their solutions. *Journal of Mathematical Physics*, 12:2116–2123, 1971.

[2] S. Diederich and M. Opper. Learning of correlated patterns in spin-glass networks by local learning rules. *Europhysics Letters*, 58:949–952, 1987.

[3] S. Grossberg. On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statitistical Physics*, 48:105–132, 1969.

[4] T. Heskes and B. Kappen. Learning processes in neural networks. *Submitted to Physical Review A*, 1991.

[5] J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79:2554–2558, 1982.

[6] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, AC-22:551–575, 1977.

[7] E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

[8] H. Ritter and K. Schulten. Convergence properties of Kohonen's topology conserving maps: fluctuations, stability, and dimension selection. *Biological Cybernetics*, 60:59–71, 1988.

[9] D. Rumelhart, G. Hinton, and R. Williams. Learning representation by back-propagating errors. *Nature*, 323:533–536, 1986.