Learning with "Relevance": Using a third factor to stabilise Hebbian learning

Running title: Using a third factor to stabilise Hebbian learning

Bernd Porr[1] and Florentin Wörgötter[2]

[1] Department of Electronics & Electrical Engineering, University of Glasgow
Glasgow, GT12 8LT, Scotland
[2] Bernstein Centre for Computational Neuroscience, University of Göttingen
Bunsenstr. 10, 37073 Göttingen, Germany

Abstract

It is a well known fact that Hebbian learning is inherently unstable because of its self-amplifying terms: the more a synapse grows the stronger the post-synaptic activity and therefore the faster the synaptic growth. This unwanted weight growth is driven by the auto-correlation term of Hebbian learning where the same synapse drives its own growth. On the other hand the cross-correlation term performs actual learning where different inputs are correlated with each other. Consequently, we would like to minimise the auto-correlation and maximise the cross-correlation. Here we show that we can achieve this with a third factor which switches on learning when the auto-correlation is minimal or zero and the cross-correlation is maximal. The biological counterpart of such a third factor is a neuromodulator which switches on learning at a certain moment in time. We show in a behavioural experiment that our three factor learning clearly outperforms classical Hebbian learning.

1 Introduction

Hebbian learning (Hebb, 1949) inherently suffers from a stability problem, which can be simply stated as: If a synapse grows the output will grow, leading to further growth of the synapse and so on. Hence in an auto-correlative manner, such a synapse influences its own growth. As long as there are only direct input-output correlations to be learned (e.g. facilitation of neuronal activity) this may not be a problem. However, there exist many cases where it is of vital importance to learn the (cross-)correlation *between* inputs. The most prominent example is classical conditioning (Pavlov, 1927; Balkenius and Morén, 1998) where the correlation between unconditioned and conditioned stimulus is learned. Also in a more technical context, when using Hebbian learning to extract the principal components of an input space, it is required to evaluate the cross-correlations while auto-correlations scale only the result (Oja, 1982; Linsker, 1988). In these and a variety of other situations the self-amplification of a Hebb-synapse may lead to a serious difficulty in the control of learning.

Here, we will concentrate on differential Hebbian learning (Kosco, 1986; Klopf, 1986; Porr and Wörgötter, 2003a) which is a variant of Hebbian learning and implements sequence learning, where two (or more) signals are correlated *in time*. In real life this can happen, for example, when heat radiation precedes a pain signal or when the vision of food precedes the pleasure of eating it. Such situations occur often during the lifetime of a creature and in these cases it is advantageous to learn reacting to the earlier stimulus, not having to wait for the later signal. Temporal sequence learning enables the animal to react to the earlier stimulus by learning an *anticipatory* action (Wörgötter and Porr, 2005).

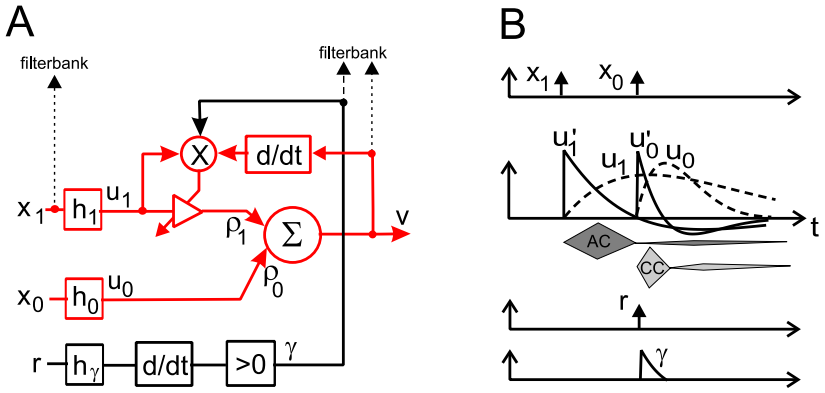

Figure 1: Learning algorithm and signal structure. **(A)** Differential Hebbian learning (red) uses the derivative of the output to control weight change. Its three-factor extension (black, solid) uses in addition a "relevance" signal r to control the timing of the learning. Dashed black lines indicate that in practical applications signals need to be fanned out into a filterbank (see below). **(B)** Signals in response to two δ-function inputs. The grey shapes (AC,CC) at the bottom denote a linear approximation of the absolute contribution of auto- and cross-correlation terms. The main part of the unwanted AC contribution comes directly after x_1. General annotations: $x_{0,1}$=input signals; r=relevance signal; $\sum$ stands for the summating neuron on which inputs converge with weights $\rho_{0,1}$. Symbols $h_{0,1,r}$ represent band-pass filters and $u_{0,1,r}$ the signals that enter the neuron; $\otimes$ denotes a correlation and the amplifier symbol stands for a changeable synaptic weight.

The auto-correlation problem can be better understood if we look at a simple neuron (see Fig. 1, red) with just two inputs u_0 and u_1. Black parts of Fig. 1 can be neglected for the time being. This neuron calculates a linear weighted sum:

$$v = \rho_0 u_0 + \rho_1 u_1 \tag{1}$$

The plasticity of the synapse ρ_1 for differential Hebbian learning (Kosco, 1986; Porr and Wörgötter, 2003a) is defined as:

$$\frac{d\rho_1}{dt} = \mu u_1 v' \tag{2}$$

where μ is the learning rate. The derivative of the post-synaptic activity implements on a phenomenological level spike timing dependent plasticity (Markram et al., 1997; Xie and

Seung, 2000; Guo-Quing and Poo, 1998; Porr and Wörgötter, 2004; Saudargiene et al., 2004) so that the order of the pre- and post-synaptic spikes determines if LTP or LTD occurs.

Now, we can substitute v' in Eq. 2 with the weighted sum of Eq. 1 and get:

$$\frac{d\rho_1}{dt} = \underbrace{\mu\rho_0 u_1 u_0'}_{cc} + \underbrace{\mu\rho_1 u_1 u_1'}_{ac} \tag{3}$$

Clearly, weight development is composed of a cross-correlation term *cc* and an auto-correlation term *ac*, which is the term which causes an unwanted weight drift: a change in the weight ρ_1 will cause a positive correlation in the auto-correlation term which in turn causes further weight change and so on.

The strategy in this paper to minimise the effect of the auto-correlation is to use the fact that in temporal sequence learning input signals happen at different moments in time. We will show that in general cross- and auto-correlation terms have little or no temporal overlap and that this will allow us to remove the unwanted auto-correlation term by using a "third factor" which switches learning on only at the moment when the auto-correlation is minimal and when the cross-correlation is maximal.

In terms of biology the application of a third factor as such is not novel. Especially in conjunction with the dopaminergic system, three factor learning has been discussed suggesting that dopaminergic responses could be related to the process of reward-based reinforcement learning (Miller et al., 1981; Schultz, 1998; Schultz and Suri, 2001). Simply this can be formalised as:

$$\frac{d}{dt}\rho = \mu \cdot \mathrm{pre}(t) \cdot \mathrm{post}(t) \cdot \mathrm{DA}(t) \tag{4}$$

were *pre* and *post* represent the pre- and post-synaptic activity at the synapse and DA is the dopamine signal.

Indeed there is experimental support in the striatum and other sub-cortical structures that Dopamine could gate the plasticity of glutamatergic synapses (for reviews see Reynolds and Wickens 2002; Wörgötter and Porr 2005). Corticostriatal synapses at medium spiny neurons will show pronounced LTP if pulsed Dopamine is present (Wickens et al., 1996). If absent, LTD arises, which is also the case for a continuous infusion of Dopamine because of D1-receptor desensitisation (Memo et al., 1982).

While many interpretations show that the dopaminergic signal is regarded as an error signal (Sutton, 1988; Mirenowicz and Schultz, 1994; Schultz et al., 1997), we suggest in this paper that it might also be used to time the learning in order to stabilise synaptic weights by minimising auto-correlation terms.

The paper is organised in the following way. In the next section we will introduce the formal framework of our three factor learning. Then we will provide a convergence proof for the open-loop condition and also demonstrate how our new learning scheme behaves in a set of standard tests. Finally we will introduce behavioural feedback (closed-loop condition) and demonstrate its stability with a simple food retrieval task.

2 ISO3-learning

We call our learning rule ISO3-learning because it is related to our differential Hebbian ISO-learning rule (Porr and Wörgötter, 2003a) where we have added a third factor.

We define the inputs to the system as x_0 (late) and x_1 (early). In all realistic situations the interval T between x_1 and x_0 is not exactly known. To account for this we introduce a filter-bank h_j at the input x_1, defining:

$$u_0 = h_0 * x_0 \tag{5}$$
$$u_j = h_j * x_1, \qquad j > 0 \tag{6}$$

with filters h_j which are given as:

$$h_j(t) = \frac{e^{-a_j t} - e^{-b_j t}}{\eta_j} \tag{7}$$

where a_j and b_j are constants defining the rise- and decay times and η_j is a normalisation constant which can be used to weight the contributions of the individual filters in a filterbank[1].

The output is a weighted sum of the filtered signals:

$$v = \rho_0 u_k + \sum_{k=1}^{N} \rho_k u_k \tag{8}$$

Now we can define ISO3-learning by (Fig. 1 A, red+black parts):

$$\frac{d\rho_k}{dt} = \mu u_k v' \gamma \tag{9}$$

where μ is the learning rate, as before. Note that the original ISO-learning rule was defined as $d\rho_k/dt = \mu u_k v'$ but is now augmented by a third factor γ.

For further analysis it is useful to rewrite Eq. 9, as in the Introduction, in the following way:

$$\frac{d}{dt}\rho_k = \mu \left(\underbrace{u_k \rho_0 u_0'}_{cc_k} + \underbrace{u_k \sum_{k=1}^{N} \rho_k u_k'}_{ac_k} \right) \gamma \tag{10}$$
$$= \mu(cc_k + ac_k)\gamma \tag{11}$$

[1]Note, that these filters differ from the ones originally used in ISO-learning. This is necessary for the convergence proof below because we need real poles for the proof instead of complex conjugate ones. However, there is no substantial difference because we have always been using highly damped resonators (e.g. $Q = 0.51$) in ISO learning which can be also modelled by the difference of two exponentials. The reader who is familiar with ISO learning will notice that a resonator around $Q = 0.5$ has a damping of $e^{-2\pi f}$ so that we define the constants a_j and b_j around $a_j = 2\pi f$ to remain compatible with our definitions from ISO-learning.

where cc_k and ac_k represent cross- and auto-correlation contributions respectively. Note, the cross-correlation term $cc_k = u_k \rho_0 u_0'$ is essentially identical to the ICO-rule (Porr and Wörgötter, 2006) given by $\frac{d\rho_1}{dt} = \mu u_1 u_0'$. In some sections we will refer to the ICO-rule to compare its behaviour to ISO- and ISO3-learning.

Furthermore, we define the signal γ by:

$$\gamma = \begin{cases} \tilde{\gamma} & \text{if} \quad \tilde{\gamma} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where

$$\tilde{\gamma}(t) = \frac{d}{dt}[r(t) * h_\gamma(t)], \tag{13}$$

where we call r the *relevance signal*. The function $h_\gamma(t)$ is also implemented by Eq. 7 where the derivative turns its low pass characteristic into a high-pass characteristic. This also guarantees that the computations in the main pathway (via $x_j \rightarrow v$) and in the relevance pathway ($r \rightarrow \gamma$) undergo the same computations, namely first low-pass filtering and then calculating the derivative.

The auto-correlation term in Eq. 11 is the one which needs to be minimised by timing the learning correctly using r. To get an idea of how this could be achieved we analyse the signal structure of this circuit (Fig. 1 B) when using δ-function inputs. Signals $u_{0,1}$ are obtained by filtering the input pulses $x_{0,1}$ with band-pass filters h which create an overlap between temporally shifted inputs, necessary for the correlation in Hebbian learning. Most importantly, however, this diagram shows the different components u_1' and u_0' of which v' is composed during learning. Because we are employing sequence learning the auto- and the cross-correlation terms happen at different moments in time which immediately suggests that one should time learning by triggering r together with x_0, because the auto-correlation is then zero. Hence, we define:

$$t_r = t_{x_0} \tag{14}$$

Using these definitions the relevance signal starts at the moment x_0 is triggered and then slowly decays. The derivative in Eq. 13 is used to eliminate the time lag obtained by the convolution of $r * h_r$ and we use only positive contributions to assure that Hebbian learning does not spuriously turn into anti-Hebbian learning.

2.1 Formal open-loop convergence condition

In order to prove that ISO3-learning with an appropriately timed r-signal can eliminate the contributions of auto-correlation terms, we will use δ-functions for all input signals. More complex input signals can be decomposed into a train of delta functions as long as the system is linear which we assume in the following derivations.

$$x_0 = a_0 \, \delta(t - T) \tag{15}$$

$$x_1 = a_1 \, \delta(t) \tag{16}$$

$$\gamma = \delta(t - T) \tag{17}$$

where x_0 and x_1 are scaled by the amplitude factors a_0 and a_1 respectively which can have any non-zero value. Having scaling factors for x_0 and x_1 and not for the relevance signal r stresses the fundamental difference between these signals: while x_0 is an error signal which can have different polarities and fluctuating amplitudes, the relevance signal always has the same amplitude and is always triggered at the moment x_0 is excited.

The idea here is to show that the distribution of weights associated with a filterbank will reach its first maximum exactly when x_0 (or r) occurs leading to a zero derivative. A situation like this has been constructed in Fig. 1 B with just a single filter, where the derivative curve reaches its maximum precisely when a delta pulse at x_0 occurs. We will now show that learning with δ-pulse inputs with a filter bank will *generate* a maximum at the moment the relevance signal is triggered and that this renders the auto-correlation term to zero.

First we have to calculate the overall weight change for ISO3-learning. The overall weight change for ρ_k is given as:

$$\Delta\rho_k = \mu \int_0^\infty u_k v' \gamma dt \tag{18}$$

by integrating the ISO3 learning rule Eq. 9 over the whole time span. This integral can also be split up into a cross- and auto-correlation term so that we get:

$$\Delta\rho_k = \underbrace{\mu \int_0^\infty u_k \rho_0 u_0' \gamma \; dt}_{\overline{cc_k}} + \underbrace{\mu \int_0^\infty u_k \sum_{j=1}^{N} \rho_j u_j' \gamma \; dt}_{\overline{ac_k}} \tag{19}$$

The integral can be solved by recalling that we have defined our signal γ as a delta function which switches learning on at time T:

$$\begin{aligned} \Delta\rho_k &= \mu \underbrace{\rho_0 u_0(0)' u_k(T)}_{cc_k} + \mu \underbrace{\left(\overbrace{\sum_j \rho_j u_j(T)}^{g_v(T)} \right)' u_k(T)}_{ac_k} && (20) \\ &= \mu \rho_0 u_0(0)' u_k(T) + \mu g_v(T)' u_k(T) && (21) \end{aligned}$$

which means that we have weight change only at time T.

The second step now is to show that at this time T the auto-correlation term ac_k remains at zero. As introduced above, this is the case if the signal $g_v(t)$ has a maximum at T so that its derivative becomes zero. The signal $g_v(t)$ is generated by the weighted filterbank responses $\rho_1 h_1, \ldots, \rho_N h_N$. Consequently, we have to show that the weights ρ_j are learned in a way that they generate a maximum at time T. At the very beginning of learning only the cross-correlation cc_k contributes to weight change because output v is still zero. Thus, for the first weight change we can concentrate on the cross-correlation term. We hope that this cross-correlation term generates a weight distribution which creates a

maximum at T so that further weight growth is only driven by the cross-correlation. Thus, we must prove that the cross-correlation term cc_k creates a maximum of g_v at T.

We note that the weight change $\Delta\rho_k$ is proportional to ρ_0, $u_0'(0)$ and $u_k'(T)$, where ρ_0 is a constant. The second term $u_0'(0)$ is the *same* for all weights so that it cannot generate a *distribution* of different weights. The only term which can change individual weights is the filter response $u_k(T)$. This means that the weight distribution must be of the form $\rho_k \propto u_k(T)$ which results in:

$$g_v(t) \propto \sum_{k=1}^{N} u_k(T)u_k(t) \tag{22}$$

We have to show that a weighted sum of filters which uses weights as their own values at time T creates a maximum at time T. This will only be possible ultimately with an infinite number of filters so that all possible T are covered:

$$g_v(t) \propto \int_0^{\infty} u_\sigma(T)u_\sigma(t)\, d\sigma \tag{23}$$

where σ scales the timing of the filters which are defined as:

$$u_\sigma(t) = \frac{e^{-ta\sigma} - e^{-tb\sigma}}{\eta_\sigma} \tag{24}$$

with a given rise- (a) and decay-time (b).

We will now solve the integral Eq. 23 with the normalisation

$$\eta_\sigma = \sqrt{\sigma(b-a)} \tag{25}$$

which guarantees that the maximum of the filterbank indeed appears at $t = T$. Substituting Eq. 24 into Eq. 23 gives:

$$g_v(t) \propto \int_{\epsilon>0}^{\infty} \frac{(e^{-ta\sigma} - e^{-tb\sigma})(e^{-Ta\sigma} - e^{-Tb\sigma})}{\sigma(b-a)} d\sigma \tag{26}$$

where ϵ is infinitely small but non-zero to avoid a singularity in the integral. To find the maximum of $g_v(t)$ we have to find the values of t where the derivative

$$g_v(t)' \propto \frac{d}{dt} \int_{\epsilon>0}^{\infty} \frac{(e^{-ta\sigma} - e^{-tb\sigma})(e^{-Ta\sigma} - e^{-Tb\sigma})}{\sigma(b-a)} d\sigma \tag{27}$$

becomes zero. This is an exponential integral which can be solved by exchanging differentiation and integration. With that trick the σ in the denominator vanishes which makes the successive integration possible. We refer the reader to the appendix where we derive the solution step-by-step. Here, we show directly the result:

$$g_v(t)' \propto \frac{1}{a-b}\left(\frac{-ae^{-\epsilon(at+bT)}}{at+bT} - \frac{-ae^{-\epsilon a(t+T)}}{a(t+T)} + \frac{-be^{-\epsilon(aT+bt)}}{aT+bt} - \frac{-be^{-\epsilon b(t+T)}}{b(t+T)}\right) \tag{28}$$

For small numbers of $\epsilon \to 0$ the exponentials in the numerator converge towards one, which yields:

$$g_v(t)' \propto \frac{T(t-T)(a-b)}{(at+bT)(aT+bt)(t+T)} \tag{29}$$

This term becomes zero for $t = T$ which is the desired result: the derivative of the filterbank is zero at $t = T$ so that the auto-correlation is zero at the moment the relevance signal r is triggered.

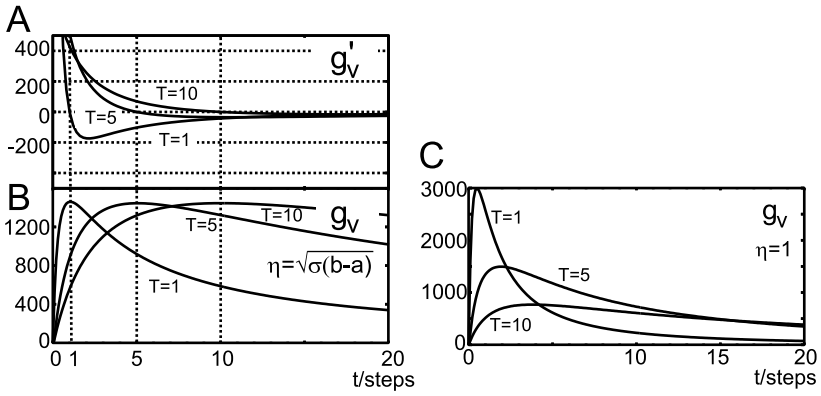

Figure 2: **A)** Plot of Eq. 28 for different values of $T = 1, 5, 10$. $a = 0.0001$, $b = 0.0005$, $c = 1$ and $\eta_\sigma = \sqrt{\sigma(b-a)}$ and **B)** plot of Eq. 26 with the same parameters. **C)** Filters have been normalised with $\eta = 1$ instead.

Fig 2A shows a plot of Eq. 28 for different values of T. The choice of a and b is not critical as long as they are not identical. Here we have set the constants a and b to small values so that the integration takes into account slow rise- and decay times. It is clear that the extremum is at the desired position $t = T$.

The integral Eq 26 has no closed form solution, but can be integrated numerically, where the results are shown in Fig 2B. We have chosen $T = 1, 5$ and $T = 10$ as the time between x_1 and x_0, γ.

But also with different, "wrong" normalisations we get interesting properties as shown for the normalisation $\eta = 1$ where we get a maximum which is at about half of T (see Fig 2C). This might be useful in applications where the auto-correlation term only has to be minimised but where a fast reaction is required. On the other hand with a stronger normalisation (e.g. $\eta_\sigma = \sigma(b-a)$) the maximum appears at $t > T$. Thus, with different normalisations we can fine tune the responsiveness of the system.

3 Analysing the ISO3-rule in an open-loop condition

In the following section we present two open-loop tests for our learning rule. In these tests (Figs. 3 and 4) pulse pairs have been repeatedly presented at inputs x, which converge with initial weights $\rho_0 = 1$ and $\rho_1 = 0$ at the learning unit.

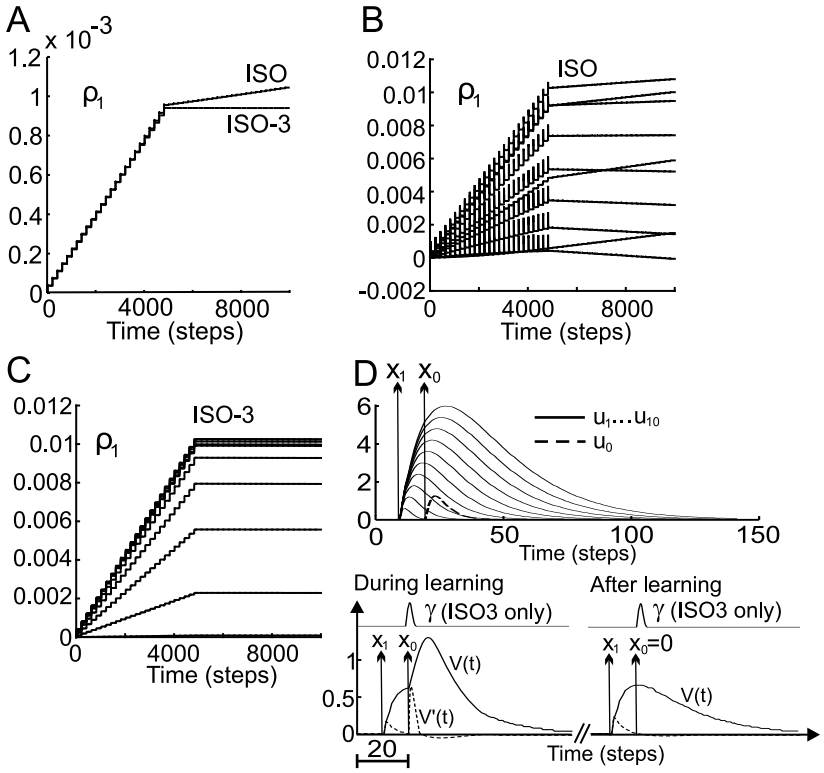

Figure 3: Comparison of the simulation results using ISO and ISO3 learning rules with a single filter **(A)** or a filter bank **(B,C)**. Experiments were performed presenting pulse-pairs as inputs. Time difference between x_1 and x_0 was $T = 10$ (x_1 always precedes x_0). At the time step 5000, x_0 was switched off. **A** Filters given by $a = 0.9\frac{2\pi}{10}$ and $b = \frac{2\pi}{10}$ were used to filter inputs x_0, x_1 and also the relevance signal r. Learning rate was $\mu = 0.005$ for the ISO learning rule and $\mu = 0.07$ for the ISO3 rule. **(B,C)** Results when using a filter bank with ten filters for signal x_1 given by $a = 0.9\frac{2\pi}{10j}, b = \frac{2\pi}{10j}, j = 1...10$. Filters with $a = 0.9\frac{2\pi}{20}, b = \frac{2\pi}{20}$ were used to filter signals x_0 and r. Learning rate $\mu = 0.001$ was used for ISO learning and $\mu = 0.002$ for ISO3. **D** shows the signals of the filter-bank u_j, and the output signal $v(t)$ when x_1 and x_0 are active and when only x_1 is active.

3.1 Comparing ISO3-learning with ISO-learning

Fig. 3 shows results for the standard test (see e.g. Porr and Wörgötter 2003a) for ISO- and ISO3 learning. Here the signal x_0 was also used to trigger the relevance signal r. Learning rates have been adjusted to produce equally strong learning for ISO and ISO3. Note, this requires larger values for μ for ISO3 than for ISO, because weight integration (Eq. 18) is limited to the surface under the small γ signal in ISO3 while it covers a broader surface in ISO. At time step 5000 the input x_0 was switched off. The corresponding signals of the filter-banks, the output during learning (x_1 and x_0 active) and after learning (x_0 switched off) are shown in panel D, respectively. According to the theory, as described in detail in Porr and Wörgötter (2003a), this should lead to weight stabilisation at ρ_1. Panel A,

however, demonstrates that weights will continue to grow for ISO after switching x_0 off. This is due to the auto-correlation influence only. The same thing happens when using a filter bank in ISO learning (Fig. 3 B), where some weights are also shrinking. Using a relevance signal prevents this unwanted effect entirely and likewise for the filter bank (Fig. 3 C). All weights become stable after x_0 has been switched off. This is due to the afore mentioned fact that v reaches its first maximum at t_{x_0} and thereby learning uses the cross-correlation term only.

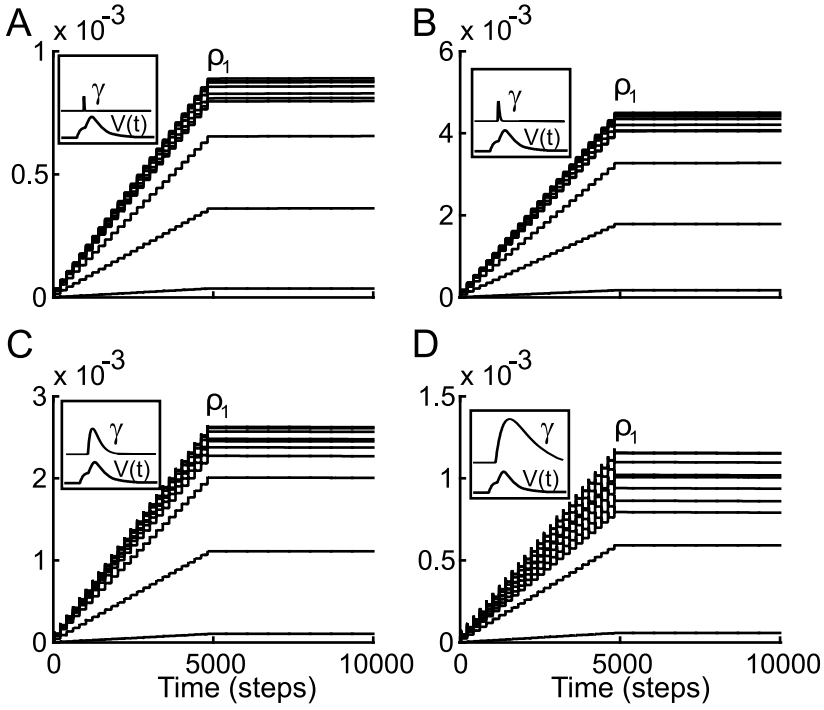

Figure 4: Simulation results using the ISO3 learning rule with *other* filters and also with varying duration a, b. Pulse pairing protocol as in Fig. 3. All filters in the signal pathways x took the form of an α-function. For the relevance pathway we used our conventional filters (Eq. 7). For x_1 we used a filter bank, setting $\alpha_{1,j} = 0.5/j, j = 1...10$, and for x_0 we set $\alpha_0 = 0.25$. Panels A-D show results for varying the shape of the filter in the relevance pathway for $\mu = 0.001$ and $t_r = t_{x_0}$ throughout. **(A)** $a = 0.9\frac{2\pi}{10}, b = \frac{2\pi}{10}$, **(B)** $a = 0.9\frac{2\pi}{20}, b = \frac{2\pi}{20}$, **(C)** $a = 0.9\frac{2\pi}{100}, b = \frac{2\pi}{100}$, **(D)** $a = 0.9\frac{2\pi}{200}, b = \frac{2\pi}{200}$.
The insets show the filtered relevance signals γ together with the output $v(t)$ which have 200 time steps on the x axis and a range of $0 \dots 1.2$ for $v(t)$ and $0 \dots 12$ for gamma.

3.2 Changing the duration of the relevance signal

In this section we test how a longer lasting relevance signal influences stability and we demonstrate that one can use different types of filters in the filter bank. To obtain the

results shown in Fig. 4 we have used α-functions

$$h(t) = te^{-\alpha t} \tag{30}$$

instead of the filters defined by Eq. 7. It is apparent from Fig. 4 that the weights stabilise as soon as the input x_0 has been switched off. Fig. 4 also shows that stability is insensitive to the length of the r-signal as long as r sets in at the same time as x_0. Varying the duration for more than one order of magnitude does not affect stability.

These findings suggest that there is a class of different filter functions for which ISO3 converges. The common feature of the filters used so far is their low pass component which generates a distinctive maximum at a certain moment in time. Together these filters are able to create a weighed sum which has its maximum at the moment the relevance signal is triggered. This means that we can choose different types of low pass filters to minimise the contribution of the autocorrelation term. This is a useful property, because the choice of the filter functions will determine how the output v is shaped. Different applications may require different types of outputs and it is now, in principle, possible to obtain them by the correct choice of filters in the filter bank (which is also true for ICO learning; Porr and Wörgötter 2006).

3.3 Summary of the result from open-loop analysis

For ISO3 we find three possible ways to stop weight growth where the third condition is the most important one. The weights stabilise:

1. trivially when $x_1 = 0$. This is obvious because then its own input is lacking.

2. when $T = 0$ or $T \to \infty$. These conditions reflect the fact the ISO3 is a differential Hebbian learning rule, related to spike timing-dependent plasticity (STDP, Saudargiene et al. (2004)), where LTP turns into LTD at $T = 0$, or where no learning takes place at large temporal intervals.

3. when $x_0 = 0$. This is the non-trivial case which has been made possible with the help of the third factor γ. As will be shown below this condition allows stable behavioural learning: as soon as the learned behaviour is able to eliminate the x_0 signal, the weights $\rho_j, j > 0$ will stop changing. This property was known and used in the original ISO learning (Porr and Wörgötter, 2003a), but weight stability could only be proven for small learning rates $\mu \to 0$, which led to the divergence of ISO learning for high learning rates. The introduction of the relevance signal r in ISO3 finally leads to the desired stability for $x_0 = 0$ also for higher learning rates.

4 Applying ISO3 in a behavioural closed loop

In the following section we will compare ISO3 with ISO in a closed-loop scenario. First we will formalise the closed loop and provide the outline of a convergence proof. Formal

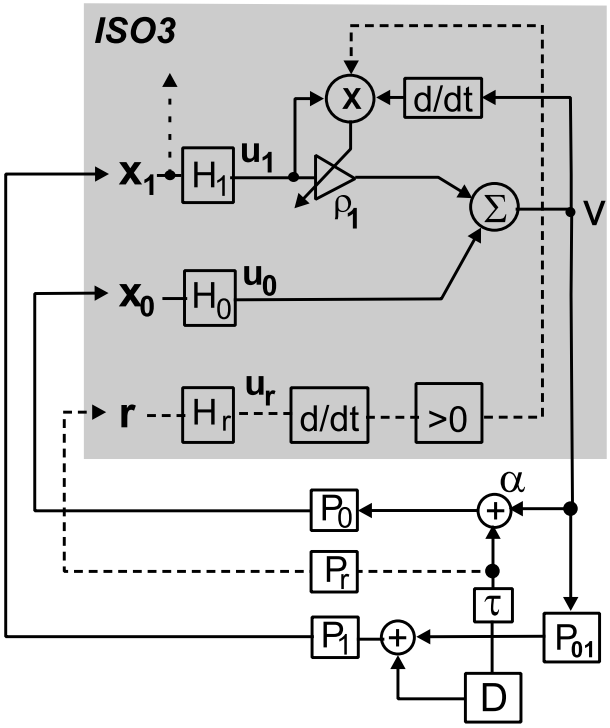

Figure 5: ISO3 embedded in a closed-loop framework. The grey box represents the learning agent, everything outside represents the environment. Most symbols as in Fig. 1. Upper case letters denote that we are treating such systems in the Laplace domain. Symbols P represent environmental transfer functions, τ is a delay, D a disturbance.

convergence proofs have been given for ISO in the limit of $\mu \to 0$ (Porr et al., 2003) and for ICO (Porr and Wörgötter, 2006) and here we will use the same arguments while taking into account the third factor.

4.1 Formalising the closed-loop situation

Fig. 5 shows how we set up our closed-loop system. This diagram is similar to the ones shown in Porr et al. (2003); Porr and Wörgötter (2006). Upper case letters denote the fact that we are treating the system in the Laplace domain. Transfer functions P are the environmental transfer functions, which are usually unknown, but well-behaved, most often leading to a delay or to some kind of low-pass filtering. These aspects are to a great extent discussed in Porr et al. (2003).

The system is built with two loops for pathways x_0 and x_1 and with one additional path for r. As in ISO- or ICO-learning, the inner loop via x_0 represents the *primary reflex*. The goal of the reflex is to compensate disturbance D at summation node α by a pre-wired response, which is achieved by setting ρ_0 so that we have classical negative feedback. This

means that we demand that the closed-loop feedback system

$$V = De^{-sT}\frac{\rho_0 H_0 P_0}{1-\rho_0 H_0 P_0} \tag{31}$$

is stable (Phillips, 2000). In this way basic behavioural patterns are established and the system is operational and prepared to learn. The outer loop is established via x_1 which is a predictive input which has the potential to generate an anticipatory action. To model the predictive nature of the input x_1 against the input x_0 we use a delay τ which delays the disturbance D so that it reaches first x_1 and then x_0.

The goal of all these systems is to adapt the behaviour, expressed by the output v, such that the primary reflex is no longer triggered via x_0. As soon as this is achieved we get $x_0 = 0$ and ρ_1 will stop to change. In this way behavioural stability arises exactly at the same time together with synaptic stability.

Finally, the r signal needs to be discussed. As mentioned before there is a major difference between the r signal and the x_0 signal: while the x_0 signal will be eliminated and becomes zero the r signal is not influenced by the output v of the ISO3-learner. Thus, the r signal will still be triggered even after successful learning when the x_0 signal has become zero.

4.2 Closed-loop convergence of ISO3 learning

In this section we will argue that convergence in the closed loop is not substantially different from the open-loop case. Convergence is assured as long as the filterbank generates a maximum at the moment the r signal is triggered and x_0 is happening. Consequently we have to find the closed-loop description of the output signal v which is in the Laplace domain:

$$V = \underbrace{\frac{\rho_0 D e^{-sT} H_0 P_0}{1-\rho_0 H_0 P_0}}_{C(s)} + \underbrace{\sum_{k=1}^{N} \rho_k \tilde{U_k}}_{A(s)} \tag{32}$$

with

$$\tilde{U_k} = \frac{U_k}{1-\rho_0 H_0 P_0} \tag{33}$$

The functions A and C can then be transformed back into the time domain and then applied in Eq. 9.

$$\frac{d}{dt}\rho_k = (\underbrace{u_k c(t)'}_{cc_k} + \underbrace{u_k a(t)'}_{ac_k})\gamma \tag{34}$$

$$= (cc_k + ac_k)\gamma \tag{35}$$

As before, it is clear that $c(t)$ forms the cross-correlation term and $a(t)$ the auto-correlation term.

The term $a(t)$ will still reach its maximum at the moment the relevance signal r is triggered because it remains constituted by the sum of low pass filtered signals (Eq 33). New is the term $1/(1-\rho_0 H_0 P_0)$ compared to the original signal U_k which introduces in the worst case a phase shift but otherwise no substantial change as long as the term does not generate more poles. This, however, is not the case because we have demanded that the pure feedback loop (Eq. 31) is stable. Consequently we can still expect the maximum at the moment the r signal is switched on because we still have a weighted sum of low pass filtered signals.

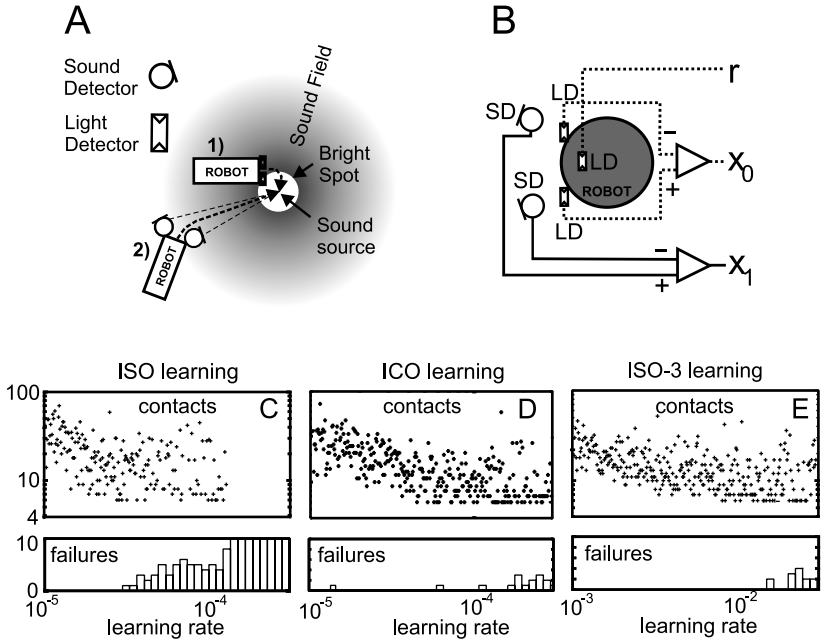

Figure 6: The robot simulation. **(A)** The robot has two pairs of sensors: light sensors which detect the food disk only in their direct proximity and sound detectors which are able to "hear" the food source from the distance. **(B)** The two light detectors (LD) establish the reflex reaction (x_0). The sound detectors (SD) establish the predictive loop (x_1). The weights $\rho_1 \dots \rho_N$ are variable and are changed either by ISO, ICO or ISO-3 learning. The signal r is generated by a third light sensor and is triggered as soon as the robot enters the food disk. The robot also has a simple retraction mechanism which operates when it collides with a wall ("retraction") which is not used for learning. The output v is the steering angle of the robot. Filters are set to $a = 0.9\frac{2\pi}{10}, b = \frac{2\pi}{10}$ for the reflex, $a = 0.9\frac{2\pi}{10k}, b = \frac{2\pi}{10k}, k = 1 \dots 5$. Reflex gain was $\rho_0 = 0.005$. **(C-E)** plot the number of contacts for three different learning rules needed for successful learning against the learning rate. In addition the number of failures against the learning rate are plotted.

5 Simulated Robot experiment

The behavioural experiment of this section has two purposes: it will give the signals x_0, x_1 and r a behavioural meaning and it will demonstrate the superiority of ISO3 compared to ISO learning. Fig. 6A,B presents the task where a simulated robot has to learn to retrieve "food disks" (Porr and Wörgötter, 2003b) which are also emitting simulated sound signals. Two sets of sensor signals are used. One sensor-type (x_0) reacts to (simulated) touch and the other sensor-type (x_1) to the sound. The reflex x_0 is established by two light detectors (LD) which draw the robot into the centre of the white disks (Fig. 6 A1). Learning must use the sound detectors (SD, Fig. 6 A2) which feed into x_1 to generate an anticipatory reaction towards the "food disk" (Verschure et al., 2003). The reflex reaction is established by the *difference* of two light dependent resistors which cause a steering reaction towards the white disk (Fig. 6B). Hence x_0 is equal to zero if both LDs are not stimulated or when they are stimulated *at the same time* which happens during a straight encounter with a disk. The latter situation occurs after successful learning. The reflex has a constant weight ρ_0 which always guarantees a stable reaction. The predictive signal x_1 is generated by using two signals coming from the sound detectors (SD). The signal is simply assumed to give the Euclidean distance from the sound source. The difference in the signals from the left and the right sound detector is a measure of the azimuth of the sound source to the robot. Successful learning leads to a turning reaction which balances both sound signals and results ideally in a straight trajectory towards the target disk ending in a head-on contact. After encountering a disk, the disk is removed and placed randomly elsewhere. Details of this experiment also show individual movement traces as shown in Porr and Wörgötter (2006), however, here we want to focus on the statistical comparison between ISO and ISO3 and try to show that ISO3 essentially performs as well as ICO, whereas ISO itself is unstable for high learning rates.

For this, we quantify successful and unsuccessful learning for increasing learning rates μ. To make the failures comparable between ISO- and ISO3-learning we have chosen the learning rates in a way that for both learning rules the contacts for successful learning are the same. Learning was considered successful when we received a sequence of five contacts with the disk at a sub-threshold value of $|x_0| < 1.1$. We recorded the actual number of contacts until this criterion was reached. The plots in Fig. 6 D-E show that fewer contacts are required for successful learning with increasing learning rates. The simulations demonstrate clearly that ISO3 learning is much more stable than the Hebbian ISO learning. It behaves very similar to ICO, for which there is no auto-correlation contribution. ISO3 learning can therefore operate at learning rates which so far have only been achieved with ICO learning but not with ISO learning.

Fig. 7A-D show how the strongest-changing weight (here ρ_9) behaves for ISO3 compared to ISO during a "food disk" experiment where we have adjusted the learning rates in such a way that weight change is similar for both ISO- and ISO3-learning. For ISO-learning there is one learning experience which leads to a correct, small weight drop close to time step 3000, but the second contact has already led to divergence. ISO3 is essentially stable, but all weights will oscillate slightly around their optimal value. As discussed above, this

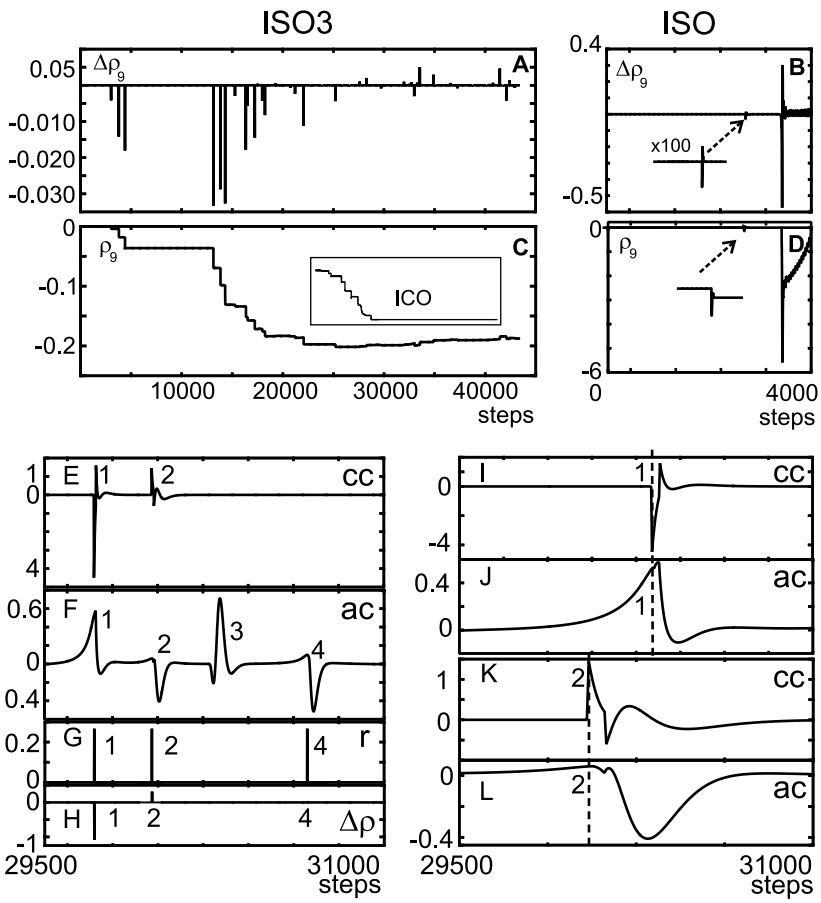

Figure 7: **A-D:** Behaviour of the strongest-growing weight ρ_9 in a "food disk" collection experiment using different learning rules. Other parameters as defined in Fig. 6. The left side shows the results from ISO3 ($\mu = 10^{-3}$) and the right side from ISO-learning ($\mu = 10^{-4}$). **(A,B)** Weight change. **(C,D)** Value of the weight ρ_9. The first learning experience ("contact") happens around time step 3000, which is magnified in panels (B,D). For ISO learning the next contact has already led to divergence. ISO3 on the other hand remains stable and ρ_9 fluctuates slightly at the end of learning. The inset in (C) shows that for ICO-learning weights will fully stabilise. **E-L:** Examples of signal shapes during four learning events ("contacts"). **(E)** Cross-correlation contribution, **(F)** auto-correlation contribution, **(G)**, the γ signal, **(H)** weight change of weight ρ_9. **(I-L)** Magnifications of events 1 and 2 from traces **(E,F)**. The dashed lines give the moment when the r signal is elicited. Note that each signal has been scaled separately.

is due to the fact that as with non-δ-function inputs the auto-correlation term cannot be fully eliminated in all cases. The remaining small fluctuation, however, will not lead to a deterioration of learning or behaviour. As long as the cross-correlation term is stronger than the auto-correlation term, weights should stabilise in the closed-loop scenario because the feedback will always correct weight drifts. To show how learning evolves without any autocorrelation term we have added inset C which shows the weight development of ICO

learning (Porr and Wörgötter, 2006) for the same food retrieval experiment.

Fig. 7E-L show in detail what the signals look like in these experiments after some learning using the same setup but with a much higher learning rate of $\mu = 0.001$. Traces (E) and (F) show the cross- (cc) and auto-correlation (ac) contributions, respectively. Due to the chosen filters cc is much shorter, but also much stronger than ac (note the different scaling). Also, it is evident that ac contributions can exist before as well as after cc. Furthermore, trace (F) shows that ac can occur without cc (events 3 and 4). Events 1,2 and 4 are associated with a relevance signal r. Trace (G) shows the corresponding r and trace (H) shows the resulting weight change. Traces (I-L) are magnifications of (E,F).

It is interesting to discuss the individual events in more detail:

1. In the first event there is a large temporal difference between the two light sensor inputs, because the robot had been approaching the food disk at an angle. This results in an early, spread-out auto-correlation term with moderate amplitude. The cross-correlation cc reaches its minimum at the moment of impact with the food disk, which is at the same moment that r is triggered. As a consequence a large negative cc contribution is summed with a much smaller positive ac contribution leading to an overall strong negative weight change. Effectively, due to the high learning rate, the system has now slightly "over-learned" the task, which becomes clear in the second event.

2. In the second event the robot also approaches the food disk at an angle but at a smaller one than in the first event above. However, the robot over-steers and touches the disk from the other side. This results in the effect that all signals are inverted and the weight is corrected upwards to a small degree. Due to the short interval when r occurs it is again obvious that the unwanted auto-correlation contribution does not enter into the weight change.

3. In the third event, the robot was directed by the predictive inputs but did not touch any food disk. Consequently, no learning should occur and the overall correlation should not deviate from zero. The auto-correlation term ac is positive and would cause an unwanted change in the weights. However, learning does not occur at event 3 because, on failing to touch, the relevance signal was not triggered at all.

4. The last event (4) shows the response when the robot approaches the food disk approximately head-on, which corresponds to $x_0 \approx 0$. Thus, the cross-correlation remains almost zero (the small existing contribution does not appear at this magnification). This shows that the robot has learned to approach the food disk from a distance and a straight trajectory towards the food disk is achieved. No weight change should happen because the learning goal $x_0 = 0$ has been reached. Learning is indeed prevented due to the fact that r occurs when, both, the cross- and the auto-correlation contributions are zero.

We have provided mathematical evidence above to illustrate that the ISO3 rule converges in a close-loop behaving system. The simulated food-disk collection experiment

shown here supports this. In particular these experiments show that the third-factor control mechanism also works with real non-δ-function inputs, for which rigorous mathematical convergence proofs are no longer possible.

In an earlier study we have shown that the auto-correlation-free ICO rule can be employed in a variety of difficult simulated and real control tasks (Porr and Wörgötter, 2006). These experiments shall not be repeated here, but the similarity of the behaviour of ISO3 in the food-disk collection supports the view that ISO3 will not demonstrate anything really new in these tasks. In addition, our simulated robot experiment also demonstrates how ISO3 input signals can be embedded in a behavioural context: the sensor signals x_0 and x_1 directly generate motor reactions and will change substantially during learning. The r-signal, however, is always triggered when the robot enters the food disk and stabilises learning by its right *timing* but not by its amplitude which always remains the same.

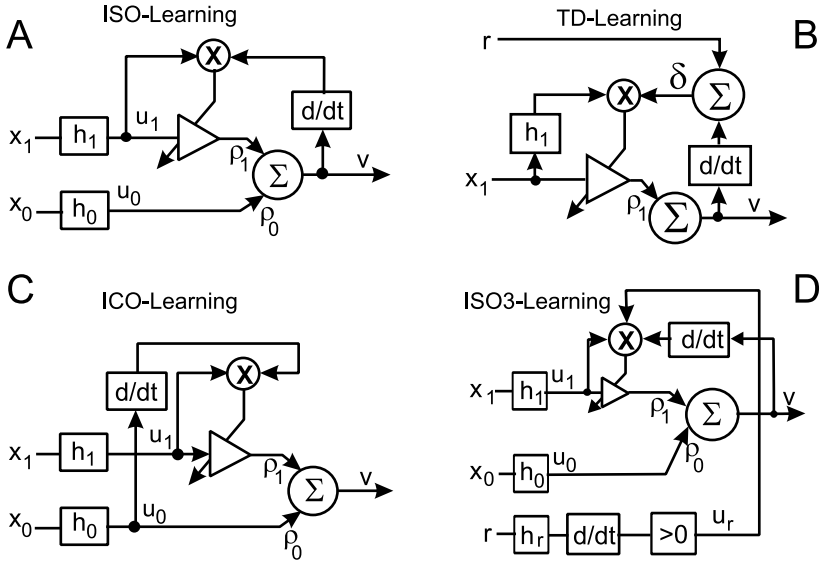

Figure 8: Four different learning rules: **(A)** ISO-learning, **(B)** TD-learning, **(C)** ICO-learning and **(D)** ISO3-learning.

6 Discussion

Correlation based temporal sequence learning dates back to the early approaches of Sutton and Barto (1981); Kosco (1986) and Klopf (1988). The design of these rules did not allow embedding them into a behavioural context and they were only treated in open-loop and mostly in conjunction with classical conditioning (for reviews see Sutton and Barto 1990; Wörgötter and Porr 2005). The success of TD-learning (Sutton, 1988) and its 'acting' extension Q-learning (Watkins and Dayan, 1992), in which both use a reward signal to control the learning, soon led to the ousting of the older correlation based approaches and they were only resurrected after they had been successfully embedded in behaving agents

(Verschure and Voegtlin, 1998; Porr and Wörgötter, 2003a). The newly introduced ISO-learning rule (re-plotted in Fig. 8 A) is successful at low but not at high learning rates because of its auto-correlation term. Porr and Wörgötter (2006) present a very simple and highly efficient solution to this problem which is shown in diagram Fig 8 C. The auto-correlation term is completely eliminated if the derivative of the output v' is replaced with the derivative of the reflex input u_0'. This rule, called *input correlation learning* (ICO), is highly stable and converges extremely fast allowing even single-shot learning as shown in several difficult real control scenarios (Porr and Wörgötter, 2006). However, it has two clear disadvantages. (1) One input, namely u_0, has now become "special". Hence learning is judged against this input and no longer against any strong driving input, which can be a problem if a subsumption architecture (Brooks, 1989) is needed where learning is driven by different inputs and not just by one input. In such subsumption architectures the driving input changes over time when one feedback loop is replaced by another which in turn leads to another driving input after learning. This argument can be reversed if we recall that ISO3 is implementing differential Hebbian learning which computes predictions: the third factor defines the moment in time when learning takes place. This provides an opportunity to self organise the development of weights which grow if their corresponding inputs can predict post-synaptic activity and which shrink if their corresponding signals are too late at the moment when the relevance signal is triggered. In such a self organised network strong inputs develop by themselves and need not be defined. This offers also new opportunities for self organised structures, for example, memory models (Durstewitz et al., 2000). (2) The second central disadvantage of ICO-learning is its low biological plausibility. ICO-learning represents a form of pure heterosynaptic plasticity, which is found only in some rare cases (Clark and Kandel, 1984; Humeau et al., 2003; Beninger and Gerdjikov, 2004; Kelley, 2004). This prompted us to search for alternative solutions to the auto-correlation problem and we have introduced the ISO3-rule for this purpose (Fig 8 D).

In the current study we have built on some earlier convergence proofs of ISO- and ICO-learning (Porr and Wörgötter, 2003a; Porr and Wörgötter, 2006) and we have focused on the problem of how to eliminate the auto-correlation term. For δ-function inputs we have now proved that eliciting the r-signal together with the later input will remove the auto-correlation term completely. In practical situations, this term cannot be fully reduced to zero, but nonetheless we found that the ISO3 rule has much better convergence properties compared to ISO. Furthermore, as in ICO also for ISO3 it is no longer necessary to use orthogonal filters as was the case for the plain ISO rule (Porr and Wörgötter, 2003a) because weight stabilisation is achieved by the generation of a maximum at x_0 and not by orthogonal filter-functions. We have shown that the maximum can also be generated by alpha functions instead of differences of exponentials. This result suggests that there is a class of functions which generate an approximately zero derivative at the moment x_0 is triggered. Looking at Eq. 22 we see that we need functions which have one maximum which can be shifted in time. If we superimpose such functions we intuitively get a maximum at the moment when the r signal is triggered. The difficult part is the right normalisation of the functions which in our case is defined by Eq. 25. Usually the difference of exponentials is normalised without the square root ($\eta = \sigma(b-a)$) which normalises the amplitude of the

functions Eq. 7 to one. However, here we have to normalise the learning (Eq. 23) which gives us the filter response two times: the filter itself and its value at the moment when r is triggered. This is a general recipe for the design of new filter functions: we need a normalisation which normalises learning (Eq. 23) instead of the functions themselves and we need filter functions which have one maximum which can be shifted in time. Such functions could be damped exponentials, alpha functions or higher order functions of the form $t^n e^{-\alpha t}$.

The relation of correlation based learning to reward-based reinforcement learning has been discussed at great length in Wörgötter and Porr (2005). Here we would like to point out one interesting novel aspect of ISO3: this rule uses only the *timing* of the r-signal to control learning. This is different from TD-learning, where a prediction error is generated which directly influences the weight values (Sutton, 1988). In other words: while the third factor in ISO3 learning determines *when* learning should happen, in RL the third factor determines *what* is learned. In machine learning the error signal is used to control value propagation in a rigorous *quantitative* way distinguishing between differently rewarding situations. The quantitative value, which an individual associates with different "rewards", is certainly also evaluated by animals and humans but it is hard to believe that the rather broad and unspecific dopaminergic signals (Fellous and Suri, 2002), which represent the majority of responses in these cell classes, would be directly used in the specific way demanded by TD-like algorithms.

Such signals seem more compatible with the assumption of 3-factor ISO3 learning, where they are only used to control the timing. It will be interesting to see if this new interpretation can be substantiated by physiological experiments in the future, for example by trying to influence the plasticity at a neuron with an ill-timed, micro-iontophoretically applied dopamine burst.

7 Acknowledgements

We thank Tomas Kulvicius and Maria Thompson for running the open-loop and closed-loop simulations, respectively. We thank Christoph Kolodziejski, David Murray Smith, Nicholas Bailey, John Williamson, John O'Reilly and Vi Romanes for their constructive feedback.

The authors acknowledge the support of the European Commission, IP-Project "PACO-PLUS" (IST-FP6-IP-027657).

A Solving the exponential integral

We have to solve the integral Eq. 26 which can be rewritten in the form:

$$g_v(t) \propto \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma(at-bT)}}{\sigma(a-b)} d\sigma - \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma a(t+T)}}{\sigma(a-b)} d\sigma + \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma(aT-bt)}}{\sigma(a-b)} d\sigma - \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma b(t+T)}}{\sigma(a-b)} d\sigma \quad (36)$$

from which we have to calculate its derivative $g_v(t)'$. Eq. 36 contains four terms which only differ by the arguments in the exponentials which we call $z(t)$. They can be solved in the following way:

$$\frac{d}{dt}\int_{\epsilon}^{\infty}\frac{e^{-\sigma z(t)}}{\sigma}d\sigma = \int_{\epsilon}^{\infty}\frac{d}{dt}\frac{e^{-\sigma z(t)}}{\sigma}d\sigma \quad (37)$$

$$= -\frac{dz(t)}{dt}\int_{\epsilon}^{\infty}e^{-\sigma z(t)}d\sigma \quad (38)$$

$$= \frac{dz(t)}{dt}\left(0-\frac{e^{-\epsilon z(t)}}{z(t)}\right) \quad (39)$$

$$= -\frac{dz(t)}{dt}\frac{e^{-\epsilon z(t)}}{z(t)} \quad (40)$$

With this result we can solve the four terms of Eq. 36. For example, the first term of Eq. 36 has the solution:

$$\frac{1}{a-b}\frac{d}{dt}\int_{\epsilon>0}^{\infty}\frac{e^{-\sigma(at+bT)}}{\sigma}d\sigma = -\frac{1}{a-b}\frac{d(at+bT)}{dt}\frac{e^{-\epsilon(at+bT)}}{at+bT} \quad (41)$$

$$= -\frac{a}{a-b}\frac{e^{-\epsilon(at+bT)}}{at+bT} \quad (42)$$

The other terms of Eq. 36 can be solved in the same way and we arrive at Eq. 28.

References

Balkenius, C. and Morén, J. (1998). Computational models of classical conditioning: A comparative study. Technical report, Lund University Cognitive Studies 62, Lund. ISSN 1101-8453.

Beninger, R. and Gerdjikov, T. (2004). The role of signaling molecules in reward-related incentive learning. *Neurotoxicity Research*, 6(1):91–104.

Brooks, R. A. (1989). How to build complete creatures rather than isolated cognitive simulators. In VanLehn, K., editor, *Architectures for Intelligence*, pages 225–239. Erlbaum, Hillsdale, NJ.

Clark, G. A. and Kandel, E. R. (1984). Branch-specific heterosynaptic facilitation in aplysia siphon sensory cells. *Proc.Natl.Acad.Sci.(USA)*, 81(8):2577–2581.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature neurosci. suppl.*, 3:1184–1191.

Fellous, J. M. and Suri, R. E. (2002). The roles of dopamine. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA, second edition edition.

Guo-Quing, B. and Poo, M.-M. (1998). Synaptic modifications in cultured hippocampus neurons. *J. Neurosci.*, 18(24):10464–10472.

Hebb, D. O. (1949). *The organization of behavior: A neurophychological study.* Wiley-Interscience, New York.

Humeau, Y., Shaban, H., Bissière, S., and Lüthi, A. (2003). Presynaptic induction of heterosynaptic associative plasticity in the mammalian brain. *Nature*, 426(6968):841–845.

Kelley, A. E. (2004). Ventral striatal control of appetitive motivation: role in ingestive behaviour and reward-related learning. *Neurosci. and Biobehav. Reviews*, 27:765–776.

Klopf, A. H. (1986). A drive-reinforcement model of single neuron function. In Denker, J. S., editor, *Neural Networks for Computing: Snowbird, Utah*, volume 151 of *AIP conference proceedings*, New York. American Institute of Physics.

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiol.*, 16(2):85–123.

Kosco, B. (1986). Differential hebbian learning. In Denker, J. S., editor, *Neural Networks for computing: Snowbird, Utah*, volume 151 of *AIP conference proceedings*, pages 277–282, New York. American Institute of Physics.

Linsker, R. (1988). Self-organisation in a perceptual network. *Computer*, 21(3):105–117.

Markram, H., Lübke, J., Frotscher, M., and Sakman, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275:213–215.

Memo, M., Lovenberg, W., and Hanbauer, I. (1982). Agonist-induced subsensitivity of adenylate cyclase coupled with a dopamine receptor in slices from rat corpus striatum. *Proc. Natl. Acad. Sci. USA*, 79:4456–4460.

Miller, J. D., Sanghera, M. K., and German, D. C. (1981). Mesencephalic dopaminergic unit activity in the behaviorally conditioned rat. *Life Sci.*, 29:1255–1263.

Mirenowicz, J. and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.*, 72(2):1024–1027.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15(3):267–273.

Pavlov, I. (1927). *Conditional Reflexes.* Oxford Univ. Press, London.

Phillips, C. L. (2000). *Feedback control systems.* Prentice-Hall International (UK), London.

Porr, B., von Ferber, C., and Wörgötter, F. (2003). ISO-learning approximates a solution to the inverse-controller problem in an unsupervised behavioural paradigm. *Neural Comp.*, 15:865–884.

Porr, B. and Wörgötter, F. (2003a). Isotropic Sequence Order learning. *Neural Comp.*, 15:831–864.

Porr, B. and Wörgötter, F. (2003b). Isotropic sequence order learning in a closed loop behavioural system. *Roy. Soc. Phil. Trans. Math., Phys. & Eng. Sciences*, 361(1811):2225–2244.

Porr, B. and Wörgötter, F. (2004). Analytical solution of spike-timing dependent plasticity based on synaptic biophysics. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16, Cambridge, MA. MIT Press.

Porr, B. and Wörgötter, F. (2006). Strongly improved stability and faster convergence of temporal sequence learning by utilising input correlations only. *Neural Comp.*, 18(6):1380–1412.

Reynolds, J. N. and Wickens, J. R. (2002). Dopamine dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507–521.

Saudargiene, A., Porr, B., and Wörgötter, F. (2004). How the shape of pre- and postsynaptic signals can influence stdp: A biophysical model. *Neural Comp.*, 16:595–626.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80(1):1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Schultz, W. and Suri, R. E. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comp.*, 13(4):841–862.

Sutton, R. (1988). Learning to predict by method of temporal differences. *Machine Learning*, 3(1):9–44.

Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170.

Sutton, R. S. and Barto, A. (1990). Time-derivative models of Pavlovian reinforcement. In Gabriel, M. and Moore, J., editors, *Learning and Computational Neuroscience*, pages 497–537. MIT-press, Cambridge, MA.

Verschure, P. and Voegtlin, T. (1998). A bottom-up approach towards the acquisition, retention, and expression of sequential representations: Distributed adaptive control III. *Neural Networks*, 11:1531–1549.

Verschure, P. F. M. J., Voegtlin, T., and Douglas, R. J. (2003). Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, 425:620–624.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.

Wickens, J. R., Begg, A. J., and Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neurosci.*, 70:1–5.

Wörgötter, F. and Porr, B. (2005). Temporal sequence learning, prediction and control - a review of different models and their relation to biological mechanisms. *Neural Comp*, 17:245–319.

Xie, X. and Seung, S. (2000). Spike-based learning rules and stabilization of persistent neural activity. In Solla, S. A., Leen, T. K., and K.-R., M., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 199–208, Cambridge, MA. MIT Press.