LETTER **Communicated by Günther Palm**

Learning with "Relevance": Using a Third Factor to Stabilize Hebbian Learning

Bernd Porr
B.Porr@elec.gla.ac.uk
Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, GT12 8LT, Scotland

Florentin Wörgötter
worgott@bccn-goettingen.de
Bernstein Centre for Computational Neuroscience, University of Göttingen, 37073 Göttingen, Germany

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 postsynaptic activity, and therefore the faster the synaptic growth. This unwanted weight growth is driven by the autocorrelation 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 minimize the autocorrelation and maximize the cross-correlation. Here we show that we can achieve this with a third factor that switches on learning when the autocorrelation is minimal or zero and the cross-correlation is maximal. The biological counterpart of such a third factor is a neuromodulator that switches on learning at a certain moment in time. We show in a behavioral 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: if a synapse grows, the output will grow, leading to further growth of the synapse, and so on. Hence, in an autocorrelative 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 & 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

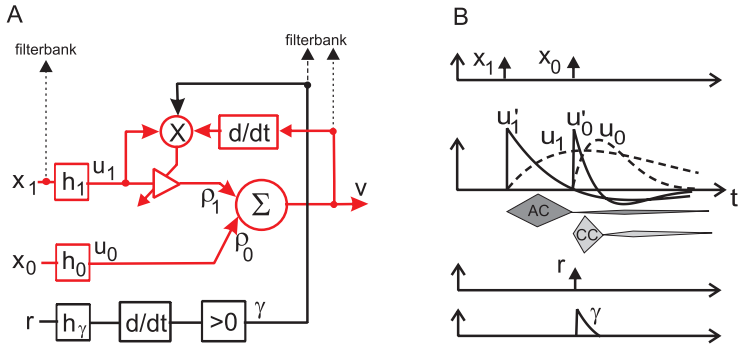

Figure 1: Learning algorithm and signal structure. (A) Differential Hebbian learning (red) uses the derivative of the output to control weight change. Its three-factor extension (black, solid line) uses in addition a "relevance" signal r to control the timing of the learning. The dashed black lines indicate that in practical applications, signals need to be fanned out into a filter bank (see below). (B) Signals in response to two δ-function inputs. The gray 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 bandpass 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.

is required to evaluate the cross-correlations, while autocorrelations 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 serious difficulty in the control of learning.

Here, we concentrate on differential Hebbian learning (Kosco, 1986; Klopf, 1986; Porr & 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 & Porr, 2005).

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

$$v = \rho_0 u_0 + \rho_1 u_1. \tag{1.1}$$

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

$$\frac{d\rho_1}{dt} = \mu u_1 v', \tag{1.2}$$

where μ is the learning rate. The derivative of the postsynaptic activity implements on a phenomenological level spike-timing-dependent plasticity (Markram, Lübke, Frotscher, & Sakmann, 1997; Xie & Seung, 2000; Guo-Quing & Poo, 1998; Porr & Wörgötter, 2004; Saudargiene, Porr, & Wörgötter, 2004) so that the order of the pre- and postsynaptic spikes determines if long-term potentiation (LTP) or long-term depression (LTD) occurs.

Now, we can substitute v' in equation 1.2 with the weighted sum of equation 1.1 and get

$$\frac{d\rho_1}{dt} = \underbrace{\mu\rho_0 u_1 u_0'}_{cc} + \underbrace{\mu\rho_1 u_1 u_1'}_{ac}. \tag{1.3}$$

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

The strategy in this letter to minimize the effect of the autocorrelation 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 autocorrelation terms have little or no temporal overlap and that this will allow us to remove the unwanted autocorrelation term by using a third factor that switches learning on only at the moment when the autocorrelation is minimal and 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, Sanghera, & German, 1981; Schultz, 1998; Schultz & Suri, 2001). Simply this can be formalized as

$$\frac{d}{dt}\rho = \mu \cdot \mathrm{pre}(t) \cdot \mathrm{post}(t) \cdot \mathrm{DA}(t), \tag{1.4}$$

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

Indeed there is experimental support in the striatum and other subcortical structures that dopamine could gate the plasticity of glutamatergic

synapses (for reviews, see Reynolds & Wickens, 2002; Wörgötter & Porr, 2005). Corticostriatal synapses at medium spiny neurons will show pronounced LTP if pulsed dopamine is present (Wickens, Begg, & Arbuthnott, 1996). If absent, LTD arises, which is also the case for a continuous infusion of dopamine because of D1-receptor desensitization (Memo, Lovenberg, & Hanbauer, 1982).

While many interpretations show that the dopaminergic signal is regarded as an error signal (Sutton, 1988; Mirenowicz & Schultz, 1994; Schultz, Dayan, & Montague, 1997), we suggest in this letter that it might also be used to time the learning in order to stabilize synaptic weights by minimizing autocorrelation terms.

The letter is organized in the following way. In the next section, we introduce the formal framework of our three-factor learning. Then we provide a convergence proof for the open-loop condition and demonstrate how our new learning scheme behaves in a set of standard tests. Finally, we introduce behavioral 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 & 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:

$$u_0 = h_0 * x_0 \tag{2.1}$$

$$u_j = h_j * x_1, \qquad j > 0 \tag{2.2}$$

with filters h_j, which are given as

$$h_j(t) = \frac{e^{-a_j t} - e^{-b_j t}}{\eta_j}, \tag{2.3}$$

where a_j and b_j are constants defining the rise and decay times and η_j is a normalization constant that can be used to weight the contributions of the individual filters in a filter bank.[1]

[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 also be modeled by the difference of two exponentials. The reader who is familiar

The output is a weighted sum of the filtered signals,

$$v = \rho_0 u_k + \sum_{k=1}^{N} \rho_k u_k. \tag{2.4}$$

Now we can define ISO3 learning by (see Figure 1A, red and black parts),

$$\frac{d\rho_k}{dt} = \mu u_k v' \gamma, \tag{2.5}$$

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 equation. 2.5, as in the section 1, in the following way:

$$\frac{d}{dt}\rho_k = \mu \left(\underbrace{u_k \rho_0 u_0'}_{cc_k} + \underbrace{u_k \sum_{k=1}^{N} \rho_k u_k'}_{ac_k} \right) \gamma, \tag{2.6}$$

$$= \mu(cc_k + ac_k)\gamma \tag{2.7}$$

where cc_k and ac_k represent cross- and autocorrelation contributions respectively. Note that the cross-correlation term $cc_k = u_k \rho_0 u_0'$ is essentially identical to the ICO rule (Porr & 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 behavior to ISO and ISO3 learning.

Furthermore, we define the signal γ by

$$\gamma = \begin{cases} \tilde{\gamma} & \text{if} \quad \tilde{\gamma} > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.8}$$

where

$$\tilde{\gamma}(t) = \frac{d}{dt}[r(t) * h_\gamma(t)], \tag{2.9}$$

where we call r the relevance signal. The function $h_\gamma(t)$ is also implemented

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.

by equation. 2.3, 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 the relevance pathway ($r \rightarrow \gamma$) undergo the same computations, namely, first low-pass filtering and then calculating the derivative.

The autocorrelation term in equation 2.7 is the one that needs to be minimized by timing the learning correctly using r. To get an idea of how this could be achieved, we analyze the signal structure of this circuit (see Figure 1B) when using δ-function inputs. Signals $u_{0,1}$ are obtained by filtering the input pulses $x_{0,1}$ with bandpass filters h, which create an overlap between temporally shifted inputs, necessary for the correlation in Hebbian learning. Most important, 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 crosscorrelation terms happen at different moments in time, which immediately suggests that one should time learning by triggering r together with x_0 because the autocorrelation is then zero. Hence, we define

$$t_r = t_{x_0}. \tag{2.10}$$

The relevance signal starts at the moment x_0 is triggered and then slowly decays. The derivative in equation 2.9 is used to eliminate the time lag obtained by the convolution of $r * h_r$, and we use only positive contributions to ensure 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 autocorrelation 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,

$$x_0 = a_0\, \delta(t - T) \tag{2.11}$$

$$x_1 = a_1\, \delta(t) \tag{2.12}$$

$$\gamma = \delta(t - T), \tag{2.13}$$

where x_0 and x_1 are scaled by the amplitude factors a_0 and a_1, respectively, which can have any nonzero 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 that 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 filter bank 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 Figure 1B with just a single filter, where the derivative curve reaches its maximum precisely when a delta pulse at x_0 occurs. We 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 autocorrelation 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

$$\Delta\rho_k = \mu \int_0^\infty u_k v' \gamma dt \tag{2.14}$$

by integrating the ISO3 learning rule, equation 2.5, over the whole time span. This integral can also be split up into a cross- and autocorrelation term so that we get

$$\Delta\rho_k = \underbrace{\mu \int_0^\infty u_k \rho_0 u_0' \gamma \; dt}_{\overline{cc_k}} + \underbrace{\mu \int_0^\infty u_k \sum_{j=1}^{N} \rho_j u_j' \gamma \; dt}_{\overline{ac_k}}. \tag{2.15}$$

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,

$$\Delta\rho_k = \mu \underbrace{\rho_0 u_0(0)' u_k(T)}_{cc_k} + \mu \underbrace{\left(\overbrace{\sum_j \rho_j u_j(T)}^{g_v(T)} \right)' u_k(T)}_{ac_k}, \tag{2.16}$$

$$= \mu \rho_0 u_0(0)' u_k(T) + \mu g_v(T)' u_k(T) \tag{2.17}$$

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

The second step now is to show that at this time T, the autocorrelation 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 filter bank 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 that creates a maximum at T so that further weight growth is driven only 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 $\rho_0, u'_0(0)$, 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 that 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

$$g_v(t) \propto \sum_{k=1}^{N} u_k(T)u_k(t). \tag{2.18}$$

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

$$g_v(t) \propto \int_0^{\infty} u_\sigma(T)u_\sigma(t)\, d\sigma, \tag{2.19}$$

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

$$u_\sigma(t) = \frac{e^{-ta\sigma} - e^{-tb\sigma}}{\eta_\sigma} \tag{2.20}$$

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

We now solve the integral, equation 2.19, with the normalization

$$\eta_\sigma = \sqrt{\sigma(b-a)}, \tag{2.21}$$

which guarantees that the maximum of the filter bank indeed appears at $t = T$. Substituting equation 2.20 into equation 2.19 gives

$$g_v(t) \propto \int_{\epsilon>0}^{\infty} \frac{(e^{-ta\sigma} - e^{-tb\sigma})(e^{-Ta\sigma} - e^{-Tb\sigma})}{\sigma(b-a)} d\sigma, \tag{2.22}$$

where ϵ is infinitely small but nonzero 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,

$$g_v(t)' \propto \frac{d}{dt} \int_{\epsilon>0}^{\infty} \frac{(e^{-ta\sigma} - e^{-tb\sigma})(e^{-Ta\sigma} - e^{-Tb\sigma})}{\sigma(b-a)} d\sigma, \tag{2.23}$$

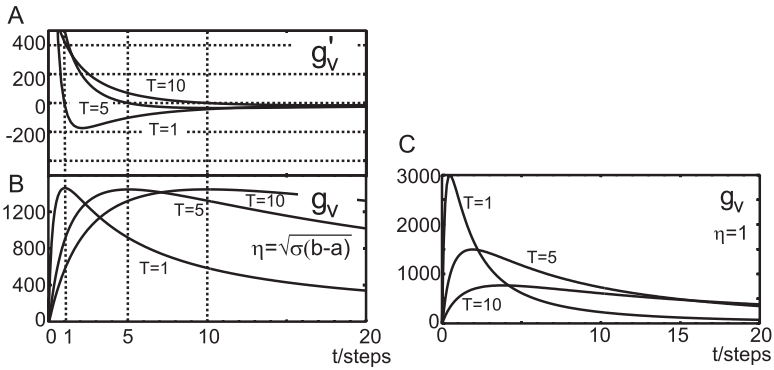

Figure 2: (A) Plot of equation 2.24 for different values of $T = 1, 5, 10$. $a = 0.0001$, $b = 0.0005$, $c = 1$, and $\eta_\sigma = \sqrt{\sigma(b-a)}$. (B) Plot of equation 2.22 with the same parameters. (C) Filters have been normalized with $\eta = 1$ instead.

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

$$g_v(t)' \propto \frac{1}{a-b}\left(\frac{-ae^{-\epsilon(at+bT)}}{at+bT} - \frac{-ae^{-\epsilon a(t+T)}}{a(t+T)} + \frac{-be^{-\epsilon(aT+bt)}}{aT+bt} - \frac{-be^{-\epsilon b(t+T)}}{b(t+T)}\right). \tag{2.24}$$

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

$$g_v(t)' \propto \frac{T(t-T)(a-b)}{(at+bT)(aT+bt)(t+T)}. \tag{2.25}$$

This term becomes zero for $t = T$, which is the desired result: the derivative of the filter bank is zero at $t = T$ so that the autocorrelation is zero at the moment the relevance signal r is triggered.

Figure 2A shows a plot of equation 2.24 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, equation 2.22, has no closed-form solution but can be integrated numerically (the results are shown in Figure 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" normalizations, we get interesting properties as shown for the normalization $\eta = 1$, where we get a maximum at about half of T (see Figure 2C). This might be useful in applications where the autocorrelation term only has to be minimized but where a fast reaction is required. With a stronger normalization (e.g., $\eta_\sigma = \sigma(b - a)$), the maximum appears at $t > T$. Thus, with different normalizations, we can fine-tune the responsiveness of the system.

3 Analyzing the ISO3 Rule in an Open-Loop Condition

In this section, we present two open-loop tests for our learning rule. In these tests (see Figures 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.

3.1 Comparing ISO3 Learning with ISO Learning. Figure 3 shows results for the standard test (see, e.g., Porr & 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 that this requires larger values for μ for ISO3 than for ISO, because weight integration (see equation 2.14) 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 Figure 3D, respectively. According to the theory, as described in detail in Porr and Wörgötter (2003a), this should lead to weight stabilization at ρ_1. Figure 3A, however, demonstrates that weights will continue to grow for ISO after switching x_0 off. This is due to the autocorrelation influence only. The same thing happens when using a filter bank in ISO learning (see Figure 3B), where some weights are also shrinking. Using a relevance signal prevents this unwanted effect entirely, and the same is true for the filter bank (see Figure 3C). All weights become stable after x_0 has been switched off. This is due to the fact that v reaches its first maximum at t_{x_0}, and thereby learning uses the cross-correlation term only.

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 Figure 4, we have used α-functions,

$$h(t) = te^{-\alpha t}, \tag{3.1}$$

instead of the filters defined by equation 2.3. It is apparent from Figure 4 that the weights stabilize as soon as the input x_0 has been switched off.

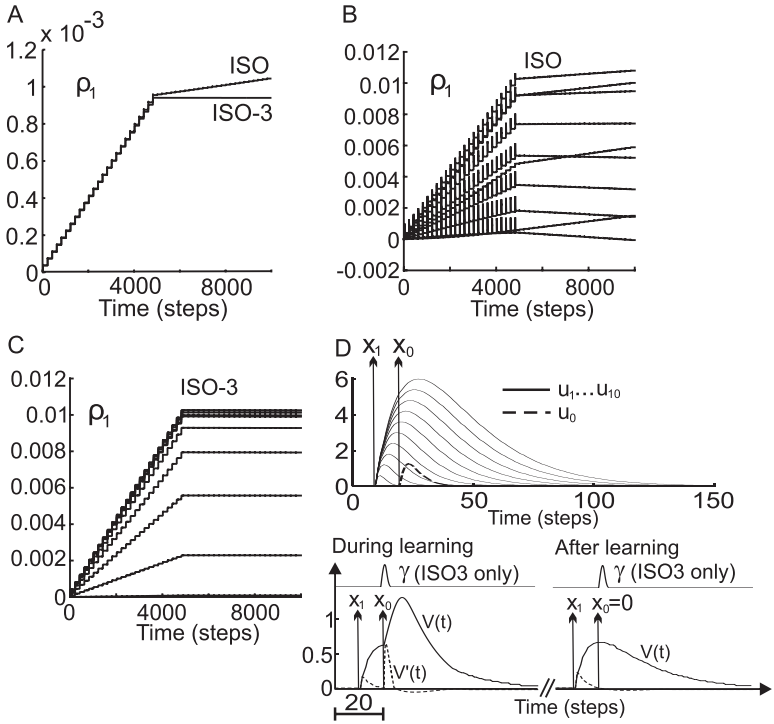

Figure 3: Comparison of the simulation results using ISO and ISO3 learning rules with a single filter (A) or a filter bank (B, C). Experiments were performed presenting pulse pairs as inputs. The time difference between x_1 and x_0 was $T = 10$ (x_1 always precedes x_0). At time step 5000, x_0 was switched 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 10 filters for signal x_1 given by $a = 0.9\frac{2\pi}{10j}, b = \frac{2\pi}{10j}$, $j = 1, \ldots, 10$. 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) 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.

Figure 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

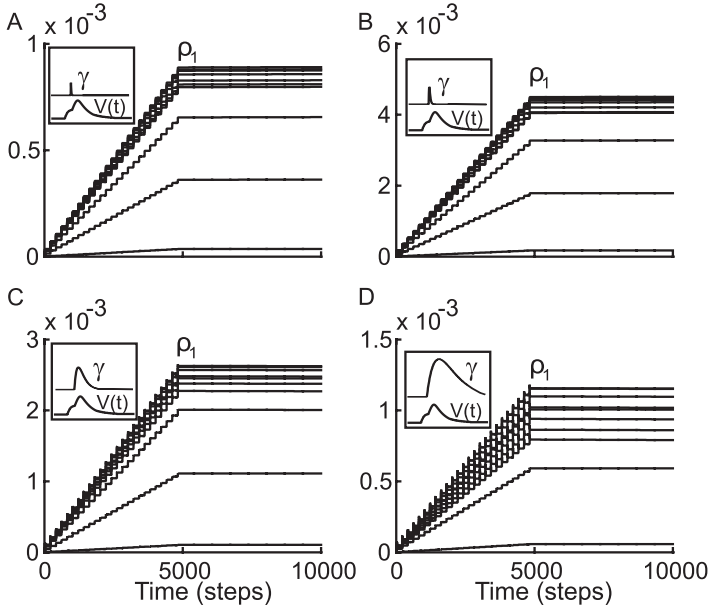

Figure 4: Simulation results using the ISO3 learning rule with other filters and also with varying duration a, b. Pulse pairing protocol as in Figure 3. All filters in the signal pathways x took the form of an α-function. For the relevance pathway, we used our conventional filters (see equation 2.3). For x_1, we used a filter bank, setting $\alpha_{1,j} = 0.5/j,\ j = 1, \ldots, 10$, and for x_0 we set $\alpha_0 = 0.25$. Panels A to 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 $0, \ldots, 1.2$ for $v(t)$ and $0, \ldots, 12$ for gamma.

that has its maximum at the moment the relevance signal is triggered. This means that we can choose different types of low-pass filters to minimize 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 & 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 stabilize:

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 that 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 nontrivial case, which has been made possible with the help of the third factor γ. As will be shown below, this condition allows stable behavioral learning: as soon as the learned behavior is able to eliminate the x_0 signal, the weights ρ_j, $j > 0$ stop changing. This property was known and used in the original ISO learning (Porr & Wörgötter, 2003a), but weight stability could be proven for only 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 Behavioral Closed Loop

In the following section, we compare ISO3 with ISO in a closed-loop scenario. First, we formalize the closed loop and provide the outline of a convergence proof. Formal convergence proofs have been given for ISO in the limit of $\mu \to 0$ (Porr, von Ferber, & Wörgötter, 2003) and for ICO (Porr & Wörgötter, 2006), and here we use the same arguments while taking into account the third factor.

4.1 Formalizing the Closed-Loop Situation. Figure 5 shows how we set up our closed-loop system. This diagram is similar to the ones shown in Porr et al. (2003) and Porr and Wörgötter (2006). Uppercase letters denote 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 prewired 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,

$$V = De^{-sT}\frac{\rho_0 H_0 P_0}{1 - \rho_0 H_0 P_0}, \tag{4.1}$$

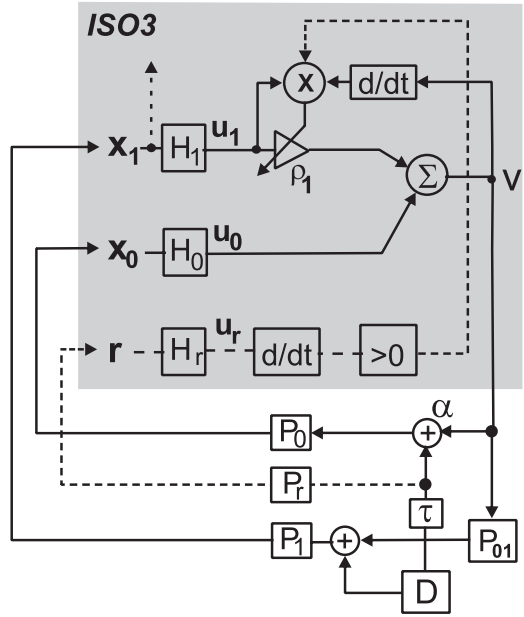

Figure 5: ISO3 embedded in a closed-loop framework. The gray box represents the learning agent; everything outside represents the environment. Most symbols are as in Figure 1. Uppercase letters denote that we are treating such systems in the Laplace domain. P represents environmental transfer functions,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

is stable (Phillips, 2000). In this way, basic behavioral patterns are established, and the system is operational and prepared to learn. The outer loop is established by x_1 which is a predictive input with 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 behavior, expressed by the output v, such that the primary reflex is no longer triggered by x_0. As soon as this is achieved, we get $x_0 = 0$, and ρ_1 will stop to change. In this way, behavioral stability arises exactly at the same time as 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 argue that convergence in the closed loop is not substantially different from the open-loop case. Convergence is ensured as long as the filter bank 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,

$$V = \underbrace{\frac{\rho_0 D e^{-sT} H_0 P_0}{1 - \rho_0 H_0 P_0}}_{C(s)} + \underbrace{\sum_{k=1}^{N} \rho_k \tilde{U}_k}_{A(s)}, \tag{4.2}$$

with

$$\tilde{U}_k = \frac{U_k}{1 - \rho_0 H_0 P_0}. \tag{4.3}$$

The functions A and C can then be transformed back into the time domain and applied in equation 2.5,

$$\frac{d}{dt}\rho_k = (\underbrace{u_k c(t)'}_{cc_k} + \underbrace{u_k a(t)'}_{ac_k})\gamma \tag{4.4}$$

$$= (cc_k + ac_k)\gamma. \tag{4.5}$$

As before, it is clear that $c(t)$ forms the cross-correlation term and $a(t)$ the autocorrelation 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 (see equation 4.3). 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, equation 4.1, 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.

5 Simulated Robot Experiment

The behavioral experiment of this section has two purposes: it will give the signals x_0, x_1, and r a behavioral meaning, and it will demonstrate the superiority of ISO3 compared to ISO learning. Figures 6A and 6B present the task where a simulated robot has to learn to retrieve "food disks" (Porr & Wörgötter, 2003b) which are also emitting simulated sound signals. Two sets

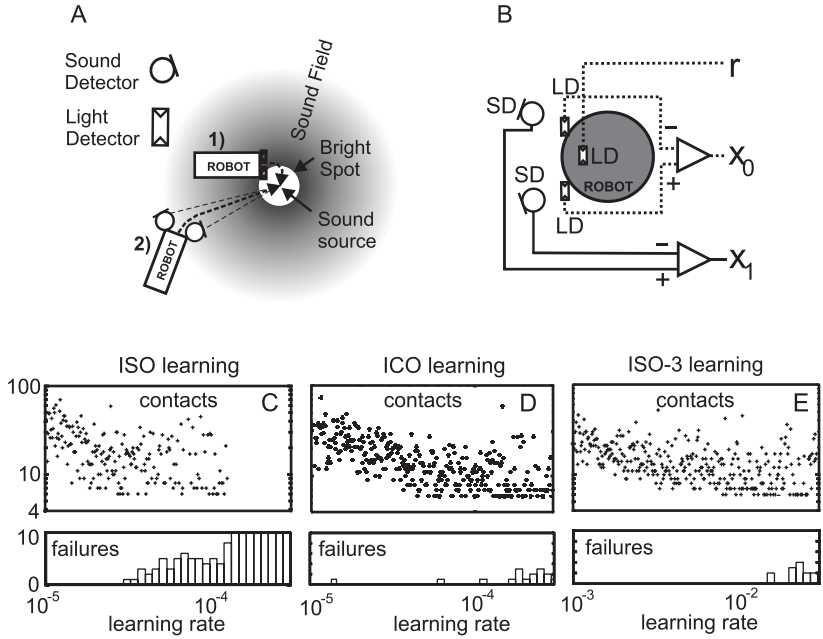

Figure 6: The robot simulation. (A) The robot has two pairs of sensors: light sensors that detect the food disk only in their direct proximity and sound detectors that are able to "hear" the food source from the distance. (B) The two light detectors (LD) establish the reflex reaction (x_0). The sound detectors (SD) establish the predictive loop (x_1). The weights $\rho_1, \ldots, \rho_N$ are variable and are changed 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 that 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, $a = 0.9\frac{2\pi}{10k}$, $b = \frac{2\pi}{10k}$, $k = 1, \ldots, 5$. Reflex gain was $\rho_0 = 0.005$. (C–E) Plots of 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.

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 center of the white disks (see Figure 6A1). Learning must use the sound detectors (SD; see Figure 6A2), which feed into x_1 to generate an anticipatory reaction toward the "food disk" (Verschure, Voegtlin, & Douglas, 2003). The reflex reaction is established by the difference of two light-dependent resistors, which cause a steering reaction toward the white disk (see Figure 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 SDs. 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 SD 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 toward the target disk, ending in a head-on contact. After a disk is encountered, 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 subthreshold value of $|x_0| < 1.1$. We recorded the actual number of contacts until this criterion was reached. The plots in Figures 6D and 6E 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 autocorrelation contribution. ISO3 learning can therefore operate at learning rates that so far have been achieved only with ICO learning, not with ISO learning.

Figures 7A to 7D 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 is due to the fact that as with non-δ-function inputs, the autocorrelation term cannot be fully eliminated in all cases. The remaining small fluctuation, however, will not lead to a deterioration of learning or behavior. As long as the cross-correlation term is stronger than the autocorrelation term, weights should stabilize 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 & Wörgötter, 2006) for the same food retrieval experiment.

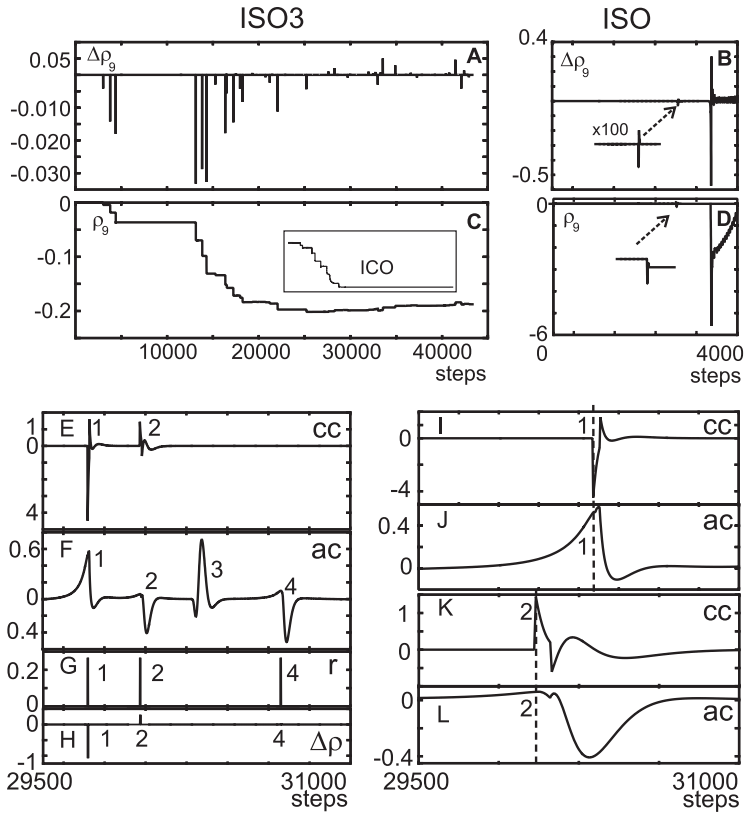

Figure 7: (A–D) Behavior of the strongest growing weight ρ_9 in a "food disk" collection experiment using different learning rules. Other parameters as defined in Figure 6. The left side shows the results from ISO3 ($\mu = 10^{-3}$) and the right side from ISO learning ($\mu = 10^{-4}$). (A, B) Weight 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 and 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 stabilize. (E–L) Examples of signal shapes during four learning events ("contacts"). (E) Cross-correlation contribution. (F) Autocorrelation contribution. (G) The γ signal. (H) Weight change of weight ρ_9. (I–L) Magnifications of events 1 and 2 from E and F. The dashed lines give the moment when the r signal is elicited. Note that each signal has been scaled separately.

Figures 7E and 7L 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 7E and 7F show the cross-(cc) and

autocorrelation (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, Figure 7F shows that ac can occur without cc (events 3 and 4). Events 1, 2, and 4 are associated with a relevance signal r. Figure 7G shows the corresponding r, and Figure 7H shows the resulting weight change. Figures 7I to 7L are magnifications of Figures 7E and 7F.

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 autocorrelation 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 "overlearned" the task, which becomes clear in the second event.
2. In the second event, the robot approaches the food disk at an angle but at a smaller one than in the first event. However, the robot oversteers and touches the disk from the other side. This results in the effect that all signals are inverted, and the weight is corrected upward to a small degree. Due to the short interval when r occurs, it is again obvious that the unwanted autocorrelation 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 autocorrelation 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 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 toward 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 autocorrelation contributions are zero.

We have provided mathematical evidence above to illustrate that the ISO3 rule converges in a closed-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 showed that the autocorrelation-free ICO rule can be employed in a variety of difficult simulated and real control tasks (Porr & Wörgötter, 2006). These experiments shall not be repeated here, but the similarity of the behavior 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 demonstrates how ISO3 input signals can be embedded in a behavioral 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 stabilizes learning by its right timing but not by its amplitude, which always remains the same.

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 behavioral context, and they were only treated in open loop and mostly in conjunction with classical conditioning (for reviews, see Sutton & Barto, 1990; Wörgötter & Porr, 2005). The success of TD learning (Sutton, 1988) and its "acting" extension Q-learning (Watkins & 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 resurrected only after they had been successfully embedded in behaving agents (Verschure & Voegtlin, 1998; Porr & Wörgötter, 2003a). The newly introduced ISO learning rule (replotted in Figure 8A) is successful at low but not at high learning rates because of its autocorrelation term. Porr and Wörgötter (2006) present a very simple and highly efficient solution to this problem, which is shown in Figure 8C. The autocorrelation 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 & Wörgötter, 2006). However, it has two clear disadvantages. First, one input, 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

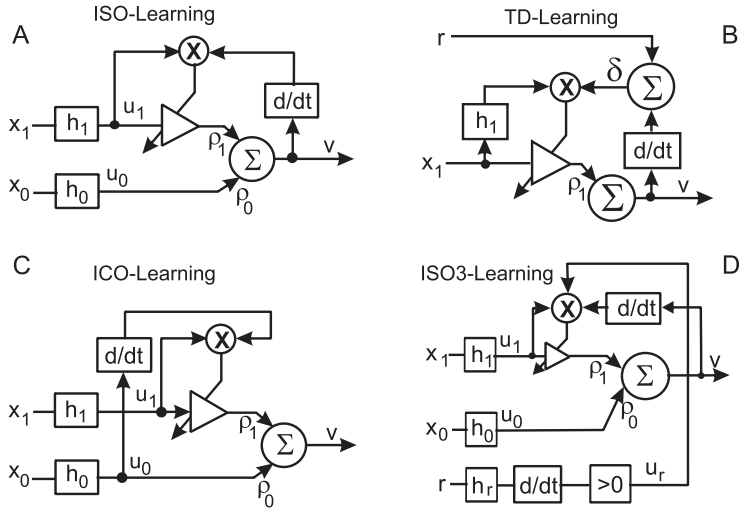

Figure 8: Four different learning rules. (A) ISO learning. (B) TD learning. (C) ICO learning. (D) ISO3 learning.

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-organize the development of weights that grow if their corresponding inputs can predict postsynaptic activity and that shrink if their corresponding signals are too late at the moment when the relevance signal is triggered. In such a self-organized network, strong inputs develop by themselves and need not be defined. This also offers new opportunities for self-organized structures, for example, memory models (Durstewitz, Seamans, & Sejnowski, 2000). 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 & Kandel, 1984; Humeau, Shaban, Bissière, & Lüthi, 2003; Beninger & Gerdjikov, 2004; Kelley, 2004). This prompted us to search for alternative solutions to the autocorrelation problem, and we have introduced the ISO3 rule for this purpose (see Figure 8D).

In this study, we have built on some earlier convergence proofs of ISO and ICO learning (Porr & Wörgötter, 2003a; Porr & Wörgötter, 2006), and we have focused on the problem of how to eliminate the autocorrelation term. For δ-function inputs, we have now proved that eliciting the r-signal together with the later input will remove the autocorrelation 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 & Wörgötter, 2003a) because weight stabilization is achieved by the generation of a maximum at x_0, 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 that generate an approximately zero derivative at the moment x_0 is triggered. Looking at equation 2.18, we see that we need functions that have one maximum that 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 normalization of the functions, which in our case is defined by equation 2.21. Usually a difference of exponentials is divided by $\eta = \sigma(b - a)$, which normalizes the amplitude (see equation 2.3) to one. However, here we have to normalize the learning (see equation 2.19), 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 normalization that normalizes learning (see equation 2.19) instead of the functions themselves, and we need filter functions with one maximum that 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 that 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 & 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 three-factor ISO3 learning, where they are used only 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.

Appendix: Solving the Exponential Integral

We have to solve the integral equation 2.22, which can be rewritten in the form

$$g_v(t) \propto \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma(at-bT)}}{\sigma(a-b)} d\sigma - \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma a(t+T)}}{\sigma(a-b)} d\sigma + \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma(aT-bt)}}{\sigma(a-b)} d\sigma - \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma b(t+T)}}{\sigma(a-b)} d\sigma, \tag{A.1}$$

from which we have to calculate its derivative $g_v(t)'$. Equation A.1 contains four terms that differ only by the arguments in the exponentials that we call $z(t)$. They can be solved in the following way:

$$\frac{d}{dt} \int_{\epsilon}^{\infty} \frac{e^{-\sigma z(t)}}{\sigma} d\sigma = \int_{\epsilon}^{\infty} \frac{d}{dt} \frac{e^{-\sigma z(t)}}{\sigma} d\sigma \tag{A.2}$$

$$= -\frac{dz(t)}{dt} \int_{\epsilon}^{\infty} e^{-\sigma z(t)} d\sigma \tag{A.3}$$

$$= \frac{dz(t)}{dt} \left(0 - \frac{e^{-\epsilon z(t)}}{z(t)}\right) \tag{A.4}$$

$$= -\frac{dz(t)}{dt} \frac{e^{-\epsilon z(t)}}{z(t)}. \tag{A.5}$$

With this result, we can solve the four terms of equation A.1. For example, the first term of equation A.1 has the solution

$$\frac{1}{a-b} \frac{d}{dt} \int_{\epsilon>0}^{\infty} \frac{e^{-\sigma(at+bT)}}{\sigma} d\sigma = -\frac{1}{a-b} \frac{d(at+bT)}{dt} \frac{e^{-\epsilon(at+bT)}}{at+bT} \tag{A.6}$$

$$= -\frac{a}{a-b} \frac{e^{-\epsilon(at+bT)}}{at+bT}. \tag{A.7}$$

The other terms of equation A.1 can be solved in the same way, and we arrive at equation 2.24.

Acknowledgments

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. We acknowledge

the support of the European Commission, IP-Project "PACO-PLUS" (IST-FP6-IP-027657), and the BMBF, BCCN-Göttingen, Proj W3.

References

Balkenius, C., & Morén, J. (1998). *Computational models of classical conditioning: A comparative study* (Tech. Rep.). Lund: Lund University.

Beninger, R., & 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 K. VanLehn (Ed.), *Architectures for intelligence* (pp. 225–239). Mahwah, NJ: Erlbaum.

Clark, G. A., & 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., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neurosci.* (Suppl.), *3*, 1184–1191.

Fellous, J. M., & Suri, R. E. (2002). The roles of dopamine. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed.). Cambridge, MA: MIT Press.

Guo-Quing, B., & 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 neuropsychological study*. New York: Wiley-Interscience.

Humeau, Y., Shaban, H., Bissière, S., & 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 J. S. Denker (Ed.), *Neural networks for computing: Snowbird, Utah*. 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 J. S. Denker (Ed.), *Neural networks for computing: Snowbird, Utah* (pp. 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., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*, 213–215.

Memo, M., Lovenberg, W., & 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., & German, D. C. (1981). Mesencephalic dopaminergic unit activity in the behaviorally conditioned rat. *Life Sci.*, *29*, 1255–1263.

Mirenowicz, J., & 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*. New York: Oxford University Press.

Phillips, C. L. (2000). *Feedback control systems*. Upper Saddle River, NJ: Prentice Hall.

Porr, B., von Ferber, C., & 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., & Wörgötter, F. (2003a). Isotropic sequence order learning. *Neural Comp.*, *15*, 831–864.

Porr, B., & 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., & Wörgötter, F. (2004). Analytical solution of spike-timing dependent plasticity based on synaptic biophysics. In S. Thrün, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems*. *16*. Cambridge, MA: MIT Press.

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

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

Saudargiene, A., Porr, B., & 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., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

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

Sutton, R. S., & Barto, A. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience* (pp. 497–537). Cambridge, MA: MIT Press.

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

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

Wickens, J. R., Begg, A. J., & 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., & 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., & Seung, S. (2000). Spike-based learning rules and stabilization of persistent neural activity. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems*, *12* (pp. 199–208). Cambridge, MA: MIT Press.

Received May 25, 2006; accepted October 9, 2006.