Biol Cybern (2008) 98:259–272
DOI 10.1007/s00422-007-0209-6

ORIGINAL PAPER

Biological Cybernetics

Mathematical properties of neuronal TD-rules and differential Hebbian learning: a comparison

Christoph Kolodziejski · Bernd Porr · Florentin Wörgötter

Received: 19 September 2007 / Accepted: 19 December 2007 / Published online: 15 January 2008

Abstract A confusingly wide variety of temporally asymmetric learning rules exists related to reinforcement learning and/or to spike-timing dependent plasticity, many of which look exceedingly similar, while displaying strongly different behavior. These rules often find their use in control tasks, for example in robotics and for this rigorous convergence and numerical stability is required. The goal of this article is to review these rules and compare them to provide a better overview over their different properties. Two main classes will be discussed: temporal difference (TD) rules and correlation based (differential hebbian) rules and some transition cases. In general we will focus on neuronal implementations with changeable synaptic weights and a time-continuous representation of activity. In a machine learning (non-neuronal) context, for TD-learning a solid mathematical theory has existed since several years. This can partly be transfered to a neuronal framework, too. On the other hand, only now a more complete theory has also emerged for differential Hebb rules. In general rules differ by their convergence conditions and their numerical stability, which can lead to very undesirable behavior, when wanting to apply them. For TD, convergence can be enforced with a certain output condition assuring that the δ-error drops on average to zero (output control). Correlation based rules, on the other hand, converge when one input drops to zero (input control). Temporally asymmetric learning rules treat situations where incoming stimuli follow each other in time. Thus, it is necessary to remember the first stimulus to be able to relate it to the later occurring second one. To this end different types of so-called *eligibility traces* are being used by these two different types of rules. This aspect leads again to different properties of TD and differential Hebbian learning as discussed here. Thus, this paper, while also presenting several novel mathematical results, is mainly meant to provide a road map through the different neuronally emulated temporal asymmetrical learning rules and their behavior to provide some guidance for possible applications.

Keywords Differential Hebbian learning · Correlation-based learning · Temporal difference learning · Reinforcement learning

C. Kolodziejski · F. Wörgötter (✉)
Bernstein Center for Computational Neuroscience,
University of Göttingen, Bunsenstr. 10, 37073 Göttingen, Germany
e-mail: worgott@bccn-goettingen.de

C. Kolodziejski
e-mail: kolo@bccn-goettingen.de

B. Porr
Department of Electronics and Electrical Engineering,
University of Glasgow, Glasgow GT12 8LT, Scotland
e-mail: b.porr@elec.gla.ac.uk

1 Introduction

Control tasks for animals or machines (systems) require sensible actions that follow from the current state of the system. As consequence of an action the state of the system may change and a new action will be elicited, and so on. This procedure describes a sequence of states and actions, which follow each other in time. Such control mechanisms can either be hardwired into the system, but many times it is more appropriate to design a learning algorithm that tries to infer the next action from the previous sequence of states. Especially in animal (or robot-) control the complexity of the world may prevent hardwiring and learning is required instead to assure enough flexibility.

In general three mechanisms can be used for learning in such situations: (1) Unsupervised learning (finding statistical structure), (2) reinforcement learning (the learner receives

a fairly unspecific reward-signal about the success of its actions) and (3) supervised learning (the learner receives a specific error signal). The distinction between these algorithmic classes is to some degree fuzzy and we will here deal with classes one and two only.

Specifically there are: reward-based reinforcement learning (Kaelbling et al. 1996; Sutton and Barto 1998 abbreviated with RL) as well as correlation based differential Hebbian learning (Kosco 1986; Klopf 1986; Porr and Wörgötter 2003), here abbreviated with CL (correlation-based learning). These algorithms differ fundamentally in their design and goal.

Most influential up to date is reinforcement learning and substantial efforts have been undertaken during the last 15 years to develop a highly successful theory of RL (Sutton 1988; Dayan 1992; Dayan and Seynowski 1994; Kaelbling et al. 1996; Sutton and Barto 1998). Mostly RL methods had been developed in conjunction with machine learning (e.g., Q-learning, Watkins 1989; Watkins and Dayan 1992) and only fewer attempts exist to design architectures which want to be more compatible with (biological) neural networks (Montague et al. 1995, 1996; Suri and Schultz 1998, 1999, 2001; Suri et al. 2001; Arleo and Gerstner 2000; Strösslin et al. 2005). When studying machine learning, it becomes clear that there is no easy way to adopt the state-action space structures used for (e.g.) Q-learning to a neural network. Most often Actor-Critic Architectures (Witten 1977; Sutton and Barto 1998) are used which emulate some aspects of the basal ganglia and the prefrontal cortex. Other biologically relevant approaches have implemented TD-learning in the context of place field guided navigation models in the hippocampus (Foster et al. 2000; Krichmar et al. 2005).

Correlation based differential Hebbian learning has been invented around 1985 (Kosco 1986; Klopf 1986, 1988) after so-called stimulus substitution models of classical (and instrumental) conditioning had been introduced (Sutton and Barto, 1981). At that time these models had not been used in any behaving closed loop system and the early CL-models had soon been superceded by the highly influential method of "temporal difference learning" (TD-learning; Sutton 1988) by which CL-methods essentially died out and became replaced by RL-approaches to which TD-learning belongs. Through TD (and related methods) it became possible to reliably learn the values of states and actions and to control the learning of an agent through reinforcement signals. Only much later CL was revived when discovering a method of how to embed CL into closed-loop control systems guaranteeing convergence of the learning process (Porr and Wörgötter 2003; Porr et al. 2003). By this CL became a possible alternative to RL. Furthermore, it has been found (Saudargiene et al. 2004; Roberts 1999) that differential Hebbian learning is related to spike-timing dependent plasticity (STDP; Gerstner et al. 1996; Markram et al. 1997).

Hence currently one finds the following situation: TD-learning is compatible to neuronal circuitry mostly at the network level, while differential hebbian learning is more directly compatible to STDP at single neurons. Only very recently several attempts have been presented in the literature to show that there is a more direct equivalence existing between these two worlds and two different proofs have been presented how RL can be emulated by a CL formalism (Florian 2007; Santiago et al. 2007; Pfister et al. 2006; Izhikevich 2007). Alas, both approaches are quite complex and the relation between RL and CL is not really straight forward.

Therefore, it may make sense to step back and first provide some simple circuits by which RL and CL can be emulated at a single unit in a time-continuous and causal way from where on it might become easier to address the question to what degree both approaches are similar. In addition, the growing number of different learning rules, especially for CL, makes it useful to provide a comparison, which may help the reader to better understand these rules when wanting to implement them.

Note, this paper deals with the different learning rules only in an open-loop context. Behavioral feedback is not treated. This would represent a second step and is beyond the scope of the current paper.

2 Basic approaches

In general we have to deal with situations where information arrives at an agent spread out over time. Hence all methods for RL and CL need to provide some kind of memory mechanism. This can be explained in the most basic way when discussing classical conditioning models (Fig. 1). In order to learn reacting to the earlier stimulus (CS), it has to be remembered in the system. To this end the concept of eligibility traces had been introduced (Hull 1939, 1943; Klopf

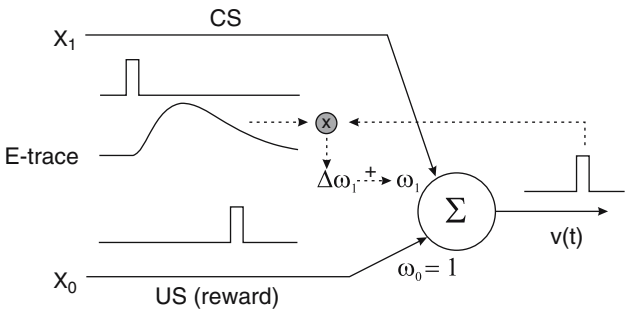

Fig. 1 Simple correlative temporal learning mechanism applying an eligibility trace (E-trace) at input x_1 to assure signal overlap at the moment when x_0 occurs; ⊗ denotes correlation. In the terminology of conditioning: *CS* conditioned stimulus, *US* unconditioned stimulus. The ω depict the synaptic weights

1972, 1982; Sutton 1988; Singh and Sutton 1996), where the synapses belonging to the earlier stimulus remain eligible for modification for some time until this trace fades. In the following we will use x as the signal input and u as the filtered input ($u = x * h$) with h being an arbitrary filter function (we will define a particular filter function later). The $*$ denotes a convolution. Furthermore, v will be the neuronal sum of the inputs weighted by the weights ω and prime will denote the derivative with respect to time: $v' = \frac{dv}{dt}$.

The first model to make use of this had been designed by Sutton and Barto (1981). In this model synaptic weights change according to

$$\frac{d\omega_1(t)}{dt} = \gamma[v(t) - \overline{v}(t)]u_1(t), \tag{1}$$

where they have introduced two eligibility traces, u_1 at the input and $\overline{v}(t)$ at the output, given by

$$u_1(t+1) = \tilde{a}u_1(t) + x_1(t) \tag{2}$$

$$\overline{v}(t+1) = \tilde{b}\overline{v}(t) + (1-\tilde{b})\, v(t), \tag{3}$$

with control parameters $\tilde{a}$ and $\tilde{b}$. Mainly they discuss the case of $\tilde{b} = 0$ where $\overline{v}(t) = v(t-1)$, which turns their rule into (Fig. 2a):

$$\textbf{S\&B}: \quad \frac{d\omega_1(t)}{dt} = \gamma[v(t) - v(t-1)]u_1(t), \tag{4}$$

Before learning this neuron will only respond to the US, while after learning it will respond to the CS as well.

TD-learning had been developed a few years later (Sutton 1988). One centrally new aspect at that point had been to introduce a reinforcement signal r, which affects the learning, but not the output of the system (Fig. 2b). Synaptic weights change according to

$$\frac{d\omega_1(t)}{dt} = [r(t) + \gamma v(t) - v(t-1)]u_1(t)$$
$$\approx [r(t) + v'(t)]u_1(t). \tag{5}$$

We define:

$$\delta(t) = r(t) + \gamma v(t) - v(t-1) \tag{6}$$

the δ-error of TD-learning, which is the mismatch between predicted (expected) and actual reward. The parameter γ is called the discount factor, which accounts for the fact that distant rewards should be valued less. For practical purposes it makes sense to also introduce a learning rate factor $\mu << 1$ in this equation, which allows controlling the speed of learning. Hence we have:

$$\textbf{TD}: \quad \frac{d\omega_1(t)}{dt} = \mu\, \delta(t)\, u_1(t) \tag{7}$$

Figure 2a and b shows the difference between the old S&B model and TD-learning. Note, in TD the reward does not enter into the output of the neuron, but only influences the learning and both diagrams are identical if we remove x_0

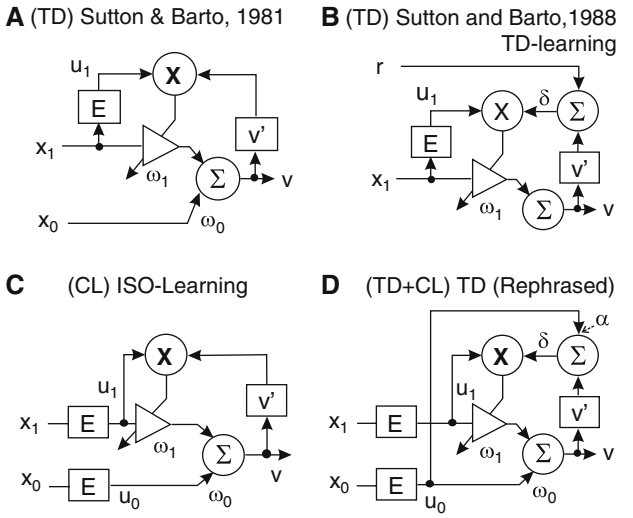

Fig. 2 Comparison of basic learning rules. Eligibility traces are denoted as E, while x_0 (or r respectively) represents the unconditioned input or reward and x_1 the conditioned input. The amplifier symbol denotes the changing synaptic weight. All models use the derivative of the postsynaptic signal in order to control the weight change. Both Sutton and Barto models **a**, **b** use eligibility traces only in the learning pathway, whereas ISO Learning **c** and TD-Rephrased **d** introduce them also in the other pathway. The rephrased TD model **d** is mixture of models **b** and **c**. We introduce a multiplicative factor α to control the impact of u_0 on the δ-value. Setting $\alpha \to 0$ we reobtain ISO Learning, setting $\omega_0 = 0$ we get TD Learning (with a filtered x_0 path instead of an unfiltered)

or r, respectively. Originally these models had been designed with the time-scale of classical or instrumental conditioning in mind, which can be seconds between CS and US. If dealing with biological systems, ultimately, however, one must match all learning models to the time scales of synaptic plasticity. When doing this, it does not make sense to separate the inputs to a neuron from their eligibility traces. Rather it seems more reasonable to assume that spikes get transformed at a membrane into several much slower processes, some of which determine the membrane potential, while others may determine plasticity. For simplicity here we might just assume *one* slow process at the input to a neuron which equally affects the neuron's output *and* its learning. When doing this we arrive at diagram Fig. 2c, which is called ISO-learning (Porr and Wörgötter 2003). The ISO-learning rule is given by

$$\textbf{ISO}: \quad \frac{d\omega_1(t)}{dt} = \mu\, v'(t)\, u_1(t) \tag{8}$$

Hence, ISO-learning is a form of differential Hebbian learning. While the main difference between the S&B and ISO rule is only that both x_1 and x_0 are filtered before they enter the neuronal summation, the ISO-learning rule behaves much different from the S&B model.

If we give up the notion of an independent reward input and, like in the old S&B model, use x_0 also as a reward, we

obtain Fig. 2d for rephrased TD-learning, where the learning rule takes the following shape:

$$\textbf{TD-r}: \quad \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} = \mu[u_0(t) + v'(t)]u_1 \tag{9}$$

Two observations can be made for this rule:

1. At the beginning of learning the output v is dominated by x_0. Hence, in this formalism, rephrased TD-learning shines up as a combination of Hebbian learning with the term $u_0(t)u_1 \approx v(t)u_1$ and differential-Hebbian learning with the term $v'(t)u_1$.
2. Also we find that this rule is (apart from a minus sign) identical to the ISO-learning rule when we set $x_0 = 0$, a case which is of great importance in the discussion on the different convergence conditions below.

3 Stability analysis for x_0, $r = 0$

In the following simulations sequences of two delta-pulses have been repetitively presented to the different systems, where x_1 is earlier than x_0 (or than r, respectively) with a constant interval $T = 30$ steps between them. The interval between pulse pairs was 300. The step by step weight change $\Delta\omega$ is calculated by integrating the respective learning rule: $\Delta\omega = \int_0^\infty \frac{\mathrm{d}\omega(t)}{\mathrm{d}t}\mathrm{d}t$. From this the development of the weights can be plotted. Learning rates μ have often been adjusted to yield similarly strong weight growth for the different systems when wanting to compare rules.

In addition we will show the different weight change curves plotting the weight change against the interval between inputs T. Note, strictly these curves reflect weight changes only for $t \to 0$, hence when w_1 is still approximate zero. For negative T the temporal order of the pulses is inverted.

The next figures (up to Fig. 7) always show how the different learning rules behave when maximally *two* eligibility traces are being used. This case applies only if one has good knowledge about the temporal difference T between the two incoming stimuli. Often, however, knowledge about T is limited, then one needs to use a set of eligibility traces for spreading out the earlier stimulus across time to make sure that at least some of these signals can be related to the later occurring x_0 or r signal. This situation is discussed later (see filter bank approach in Fig. 8).

As eligibility traces E we use in all cases a band-pass filter response defined by

$$h(t) = \frac{1}{\sigma}(e^{-at} - e^{-bt})\,\Theta(t) \tag{10}$$

with $\Theta(t)$ being the Heaviside (step) function. Actual parameters are given in the figure legends.

We are especially interested in the stability of the learning rules. All rules learn by cross-correlating two signals with each other (x_1 with x_0 or with r), correlations of x_1 with itself (auto-correlations) are normally unwanted. Hence it is of interest to subdivide the contributions of the learning rule into a cross- and an auto-correlation term by: $\Delta w = \Delta\omega^{ac} + \Delta\omega^{cc}$, the latter term drives the weight change of ω_1 during the occurrence of x_0 (or r), whereas the auto-correlation term also changes the weight in the absence of the x_0 signal. Hence, the pure auto-correlation contribution becomes visible when switching x_0 (or r) off.

This is what we do in the following diagrams by setting $x_0 = 0$ at time-step $t = 6000$ to show if the weight change for a given rule will then indeed stop. This is an important case in a closed-loop system which, for instance, has to avoid a reflex triggered by the x_0 signal (Porr and Wörgötter 2003; Porr et al. 2003; Krichmar et al. 2005).

Let us now calculate the auto- and cross-correlation contributions for the rules introduced in the previous section.

Equation 4, from the S&B 1981 model, leads to:

$$\begin{aligned}\Delta\omega_1 &= \int_0^\infty v'(t)\,u_1(t)\,\mathrm{d}t \\ &= \int_0^\infty (\omega_0 x_0(t) + \omega_1 x_1(t))'\,u_1(t)\,\mathrm{d}t \\ &= \int_0^\infty \omega_0\delta'(t-T)\,h(t)\,\mathrm{d}t + \int_0^\infty \omega_1\delta'(t)\,h(t)\,\mathrm{d}t \\ &= -\omega_0 h'(T) - \omega_1 h'(0)\end{aligned} \tag{11}$$

where we have assumed and will assume for all upcoming calculations a quasi-static[1] approach ($\omega' << 1$), exchanged $u_1(t)$ with $h(t)$ and $x_1(t)$ and $x_0(t)$ with $\delta(t)$ and $\delta(t-T)$, respectively. Additionally we used $\int_0^\infty \delta'(t-t_0)f(t) = -f'(t_0)$ (Boykina 2003).

Thus, we get:

$$\textbf{S\&B}: \quad \Delta\omega_1^{cc} = -\omega_0 h'(T) \qquad \Delta\omega_1^{ac} = -\omega_1 h'(0) \tag{12}$$

The TD-learning rule (Eq. 7) leads to similar results:

$$\textbf{TD}: \quad \Delta\omega_1^{cc} = +\omega_0 h(T) \qquad \Delta\omega_1^{ac} = -\omega_1 h'(0) \tag{13}$$

The only difference is a changed sign and the derivative of the filter h associated with the x_0 or rather r input (which is set to $\omega_0\,\delta(T)$ for comparison).

In general one finds that in all cases synapses ω_1 will grow (left panels) until the later pulse is switched off. In

[1] A quasi-static approach is commonly assumed for such calculations (Dayan and Abbott 2003) and is justified as long as μ is small. It is needed for neglecting the derivative of the weight w_1 on the right hand side of the equation and for neglecting the variability of the homogeneous solution for the calculation of the inhomogeneous part.

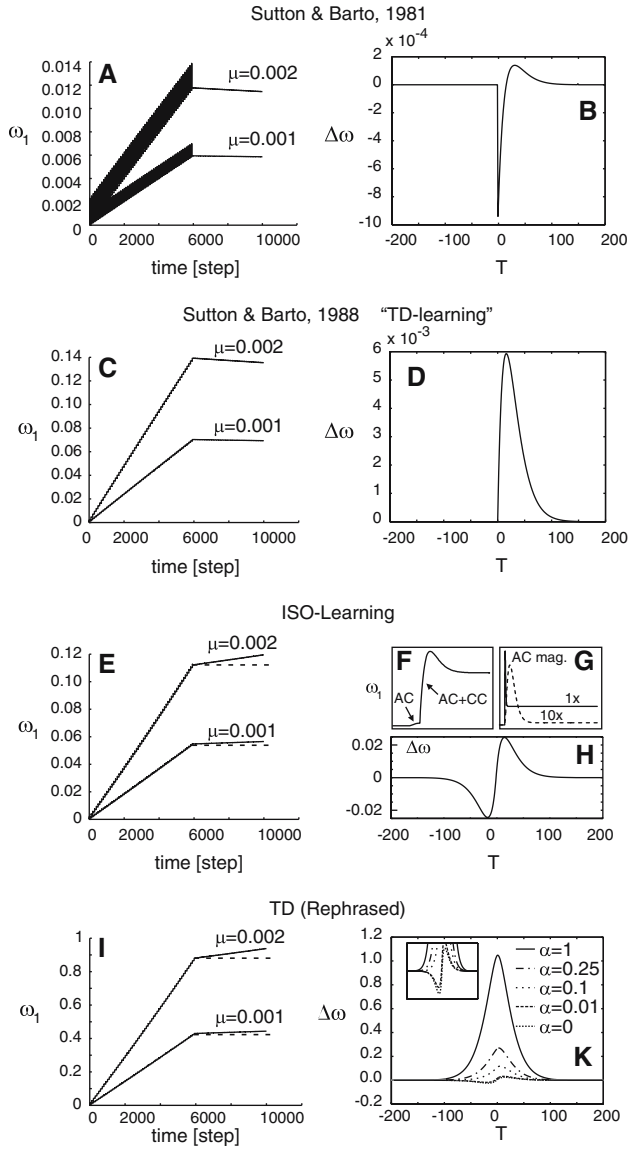

Fig. 3 Quantitative results from the architectures shown in Fig. 2. The left panels show weight growth of ω_1 during learning with two different learning rates $\mu = 0.001$ and $\mu = 0.002$. The parameters of the band-pass were $a = 0.3$, $b = 0.33$ and $\sigma = 0.03$. The right panels show the relative weight change when learning starts ($\omega_1 = 0$) for different temporal intervals T between x_0 and x_1. A positive value of T stands for $t_{x_1} < t_{x_0}$ and vice versa. **a**, **b** Model of Sutton and Barto (1981), **c**, **d** TD-learning (Sutton 1988). **e**–**h** ISO-learning (Porr and Wörgötter 2003). **f**, **g** show the temporal development of the weight change for a single input pulse pair. **f** Comparison of the AC and the combined AC + CC term. **g** Comparison of the AC term on a magnified scale for two different sampling rates (1x, 10x) to show the numerical accuracy. **i**, **k** TD-Rephrased, where **k** is plotted for different values of α and the inset shows an enlargement around the origin

the S&B model (Fig. 3a) one sees that the unfiltered input x_0 and its derivative lead to strong, needle-like excursions of the weight growth for every step, which let the line in the diagram appear broad. These disappear as soon as x_0 is switched off. In TD-learning (Fig. 3c) according to Sutton 1988, weights grow about ten times faster, which is due to the fact that the reward brings in only positive contributions. When we switch x_0 (or r) off, then we find in both cases that ω_1 drops slightly. Note, this is not what ought to be done in a TD-rule. Switching off δ would be the correct condition and obviously weights will be—by construction—stable then. Still, we want to look at the $r = 0$ case, because it directly corresponds to the $x_0 = 0$ condition of the CL-rules and shows the behavior of the pure auto-correlation contribution.

Weight drop of the S&B- and TD-rule is due to the fact that the remaining total influence on the weight growth, which now comes only from the auto-correlation term is negative. As mentioned above, for $x_0 = r = 0$, both diagrams (Fig. 2a,c) are identical (see Eqs. 12, 13) and the remaining differences in weight drop come from the different values of ω_1, reached at switch-off time.

The weight change of a single signal pair for the ISO-rule (Eq. 8) can be written as:

$$\begin{aligned} \Delta\omega_1^{cc} &= \omega_0 \int_0^\infty h(t)h'(t-T)\,\mathrm{d}t \\ &= \mathrm{sign}(T)\,\omega_0 \frac{b-a}{a+b}\frac{1}{2\sigma^2}\,h(|T|) \\ \textbf{ISO}: \qquad & \\ \Delta\omega_1^{ac} &= \omega_1 \left(\mathrm{e}^{\int_0^\infty h(t)h'(t)dt} - 1 \right) \\ &= \omega_1(\mathrm{e}^{\frac{1}{2}h^2(\infty)} - 1) = 0 \end{aligned} \tag{14}$$

where the auto-correlation term converges to zero for $t \to \infty$ as the filter function $h(t)$ possesses only one maximum and eventually decays to zero.

Additionally, we calculate the time development of the cross-correlation part to give an insight into the exact weight-change:

$$\begin{aligned} w_1^{cc}(t) = \frac{\omega_0}{\sigma^2} \Bigg(& \mathrm{sign}(T)\frac{b-a}{a+b}h(|T|) \\ & + \frac{1}{2}\left(\mathrm{e}^{(-2t+T)b} + e^{(-2t+T)a}\right) \\ & - \mathrm{e}^{-t(a+b)}\frac{ae^{aT} + be^{bT}}{a+b} \Bigg). \end{aligned} \tag{15}$$

Panels f and g of Fig. 3 show the relaxation behavior of $\Delta\omega$ for a single input pulse pair. In Fig. 3f a small early auto-correlation component (AC) is followed by a big, cross-correlation dominated hump (AC + CC) as soon as x_0 occurs. The curve relaxes to the final weight value only after some time depending on the filter characteristic of h. In Fig. 3g we magnify the auto-correlation component for a situation where we have switched x_0 off (auto-correlation only!). The dashed curve shows that, following Eq. 14, the auto-correlation indeed drops to zero after a short time. This curve was numerically calculated with a ten-fold increased sampling rate as compared to the solid-line curve shown next to it.

This represents the auto-correlation contribution, when using coarser sampling and here we see a potentially very strong source of error: The auto-correlation contribution does not vanish anymore. This is a pure numerical artifact of the integration procedure, but—as high sampling rates are often too costly (for example in real-time applications)—this artifact can strongly interfere with the convergence of ISO. Hence, we are facing two potential sources of error: (1) The tardy relaxation behavior of (essentially) the cross-correlation term (Fig. 3f). This error becomes relevant when pulse pairs follow each other in time too quickly. And (2) the non-negligible numerical error that renders the auto-correlation non zero even for long relaxation times. The other CL-rules discussed below have been invented to solve these problems.

Figure 3e shows the step-by-step behavior of ISO-learning. Weight growth is similar to that of the TD-rule. Here we find that, after switching off x_0, weights will drift upwards. This effect comes from the afore discussed numerical error and vanishes for very small integration step sizes Δt and large relaxation times t as shown by the dashed lines.

The rephrased TD-rule (Fig. 3i, Eq. 9) shows a ten-fold increase of the weight growth. This is due to the additional influence of u_0, which now also enters the learning rule. The contributions are as follows:

$$\textbf{TD-r}: \quad \begin{aligned} \Delta\omega_1^{cc} &= \omega_0 \tfrac{b-a}{a+b} \tfrac{\mathrm{sign}(T)}{2\sigma^2} h(|T|) \\ &\quad +\alpha \int_0^\infty h(t)h(t-T)dt \\ &= \tfrac{b-a}{a+b} \tfrac{\mathrm{sign}(T)}{2\sigma^2} (\omega_0 h(|T|) + \alpha H(|T|)) \\ \Delta\omega_1^{ac} &= 0 \end{aligned} \tag{16}$$

The auto-correlation term vanishes due to the same arguments as for ISO-learning. The cross-correlation term is extended by a Hebbian contribution, which causes the 10-fold increase. The antiderivative of the filter function $h(t)$ is depicted as $H(t)$. Otherwise, its behavior is very similar to that of the ISO-rule. As noted above, switching x_0 off, transforms TD-Rephrased into ISO-learning.

When looking at the weight-change curves (Fig. 3b,d,h,k), we observe that the S&B model produces strong negative weights for values of $T \approx 0$ (Fig. 3b). This effect had already been reported in their original papers (Sutton and Barto 1990).

TD-learning produces an asymmetrical weight change curve (Fig. 3d). This is a consequence of the missing E-trace in the reward pathway. If being early, the reward has already vanished before x_1 occurs and the correlation result remains zero.

The weight change curve of ISO-learning is anti-symmetrical (Porr and Wörgötter 2003). As long as both E-traces are the same, this curve will have identical lobes on both sides (Fig. 3h). This is interesting, because with this rule a completely isotropic setup can be designed, in which both synapses are allowed to change as will be discussed in Fig. 6.

The rephrased TD-rule can produce mixed properties (Fig. 3k). As such it will early during learning, as long as $\omega \approx 0$, produce essentially plain heterosynaptic Hebbian learning. This is due to the fact that the term u_0, which enters the rule, has a much stronger influence than the derivative (as discussed above), which can be neglected. Only if we introduce an artificial attenuation factor α, by which we weaken the u_0-input, we will receive a transition towards differential Hebbian learning, which will become identical to ISO-learning for $\alpha = 0$. The inset shows magnifications of the curves. This also confirms that the plain, filtered input has a much stronger influence on the learning amplitude than the derivative.

3.1 Conditions of convergence

When do all these different algorithms converge? Trivially, weight growth at ω_1 will stop as soon as $x_1 = 0$ in all cases. Also we find that for the CL-rules weights will converge if $T = 0$. Hence these systems will be essentially stable if small positive values of T are followed by small negative ones (or vice versa). Other possible convergence conditions, however, are less obvious.

For the old S&B model at that time convergence properties had not been investigated. Clearly Fig. 2 shows that for $v' = 0$ this model will converge. But this condition is not useful as it requires the output to be constant. Hence, strictly for the old S&B model convergence cannot be guaranteed. On the other hand, ISO-learning is strictly convergent for $x_0 = 0$ and weight ω_1 will then stop to change. When switching x_0 off (Fig. 3e), one can see that ω_1 is (almost) stable for small learning rates ($\mu = 0.001$). For a higher learning rate, weights drift upwards. This is the above mentioned consequence of the high numerical sensitivity of ISO learning to the integration step size.

For TD-learning there exist two different types of convergence. The first one holds for constant learning rates μ and was proven by Sutton (Sutton 1988; Dayan 1992). Here, only the expected value of the output $\langle v(t) \rangle_{t \geq t_c}$ (where t_c denotes the time convergence has been achieved) converges to an optimal value. In order to ensure convergence of the output v itself, the learning rate has to decrease over time $\mu(t) \sim f(t)$ with the condition $\sum_{t=0}^{\infty} f(t) = \infty$ and $\sum_{t=0}^{\infty} f^2(t) < \infty$ [e.g. $f(t) = 1/t$; Dayan 1992]. Since we use time independent learning rates in this article, only the mean of the output should be constant at the end of learning. Thus both types of convergence do not need a vanishing delta error δ, instead just its average over time has to be zero: $\langle \delta \rangle_{t \geq t_c} = 0$. The delta error δ itself oscillates without any reward r around zero. Simply due to the shape of the band-pass the

weight is decreasing over time and does not stabilize as mentioned before. Consequently the convergence condition for TD-learning with a constant learning rate μ is

$$0 = \langle \delta(t) \rangle_{t \geq t_c} = \langle r(t) + v'(t) \rangle_{t \geq t_c} \quad \text{or} \tag{17}$$

$$\langle v'(t) \rangle_{t \geq t_c} = - \langle r(t) \rangle_{t \geq t_c} \tag{18}$$

Note, for this condition, the output v, or rather its derivative, needs to take on a certain value as opposed to ISO-learning, where the input x_0 needs to become zero. Hence, we have one algorithm (TD-learning) where convergence is guaranteed by output-control as opposed to one other algorithm (ISO-learning), which uses input-control to guarantee convergence. Looking back at Fig. 2b, it becomes clear that setting $r = 0$ does not enforce convergence. This had been only done for auto-correlation term evaluation (comparison to the ISO-approaches).

As discussed earlier, the rephrased TD-rule shown in Fig. 2d carries properties of both algorithmic classes, convergence can be achieved for $\langle \delta \rangle = 0$ but also with $x_0 = 0$, because in this case the learning circuit is identical to the ISO-learning circuit, as already mentioned above.

Furthermore, we note that there is no generic way to rephrase the TD-specific convergence criterion $\langle \delta \rangle = 0$ into an input condition. In order to attempt this, we need to use input terms only. Let us again use TD-Rephrased to show this and define x_0 as the reward signal (Fig. 2d). Then we have $v = \omega_0 u_0 + \omega_1 u_1$. Using this in Eq. 18 we get as convergence condition for the synapse ω_1:

$$0 = u_0 + \omega_0 u_0' + \omega_1 u_1' \tag{19}$$

Here we note that the synapse ω_1, which we are supposed to stabilize, shines up on the right side. Hence, phrased in this way as an input condition, it cannot be fulfilled.

4 Features and problems of the basic architectures

Several problems exist with the above described basic approaches, most notably:

1. CL-rules are theoretically stable for $x_0 = 0$, but ISO learning is highly sensitive to numerical errors which can easily destroy convergence.
2. Input x_0 is connected to the output v in the CL-rules. Thus, CL-rules will always produce a (motor) output which can be used to generate actions.
3. By construction, TD-learning is stable for $\langle \delta \rangle = 0$.
4. As we are dealing with a single neuron with only two inputs one finds that, without additional assumptions, TD-learning cannot produce actions: The original TD-learning as depicted in Fig. 2b does not produce any output at all if starting with $\omega_1 = 0$. Also, setting $\omega_1 \neq 0$ is nonsense from a conceptional point of view. This would assume that Pawlov's dog has already some knowledge about the meaning of the bell even before the first learning experience. This, however, can be a desired aspect of TD learning as it has led to the situation that TD-learning, when considered neuronally, has only been used to *judge* the quality of but not to actually *generate* actions. Actor and Critic generically remain separate in architectures that use TD-learning to implement RL (Barto 1995). Other approaches are using additional inputs to produce a (motor) output signal (Strösslin et al. 2005). While this works, we are here concerned with a more puristic view of only using two inputs unequivocally separating Actor and Critic.

5 Modified approaches

In order to address the above stated problems it is possible to modify the existing approaches.

First we ask, can we enforce stability in CL for $x_0 = 0$ without this unwanted numerical instability? For this we need to design a learning rule for which the auto-correlation term truly vanishes.

Figure 4a shows an architecture, where we have replaced the derivative of the output in ISO-learning, with the derivative of the (later) input x_0. Hence we are only correlating inputs with each other, thus, the name of this rule: Input correlation learning (ICO, Porr and Wörgötter 2006). The learning rule is given by

$$\textbf{ICO}: \quad \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} = \mu\, u_0(t)'\, u_1(t) \tag{20}$$

The overall weight change is similar to the ISO-rule and yields:

$$\textbf{ICO}: \quad \Delta\omega_1^{cc} = \omega_0 \frac{b-a}{a+b} \frac{\mathrm{sign}(T)}{2\sigma^2} h(|T|) \quad \Delta\omega_1^{ac} \equiv 0 \tag{21}$$

Here, the auto-correlation term is by definition equal to zero.

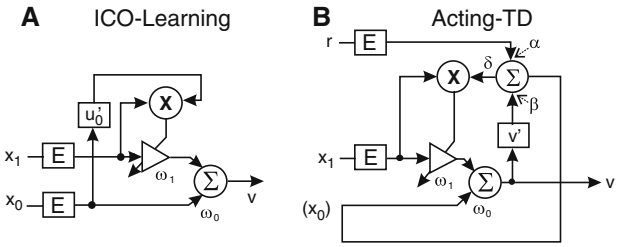

Fig. 4 Comparison of two modified learning algorithms. **a** ICO-learning. **b** Acting-TD. While Acting-TD and all architectures shown in Fig. 2 use the derivative of the postsynaptic signal, ICO-learning **a** uses the derivative of the unconditioned signal in order to control weight change. Note, x_0 in Acting-TD is identical to the delta error δ

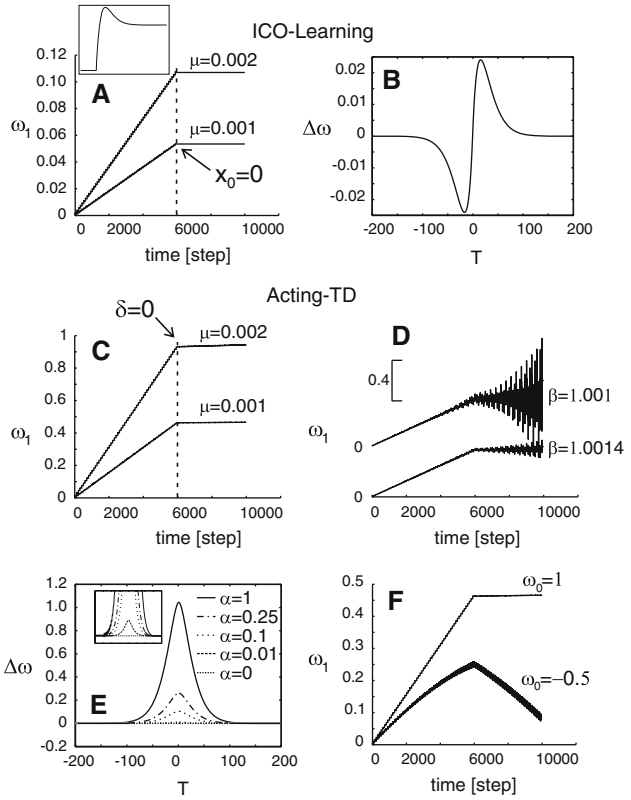

Fig. 5 Comparing the architectures shown in Fig. 4. **a**,**b** ICO-learning **d**-**f** Acting TD. Switch-off conditions ($x_0 = 0$ or $\delta = 0$) as denoted in **a**, **c**, **d**, **f**. Learning rate was $\mu = 0.001$ in **d** and **f**. The parameters of the band-pass were set to $a = 0.3$, $b = 0.33$, and $\sigma = 0.03$, throughout. The weight change for TD-Rephrased **d** is plotted for different values of α and the inset shows an enlargement around the origin

The corresponding results are shown in Fig. 5a and b. The learning window is identical to that of the ISO-rule (Fig. 3c), but now weights are stable for $x_0 = 0$. The inset in Fig. 5a shows the relaxation behavior for a single pulse-pair. In comparison to ISO learning (inset in Fig 3c) the shallow initial rising phase in missing here as there is no auto-correlation contribution. For the same reason also the "hump" is a little bit smaller. This effect, however, is barely visible even when overlaying the curves. Incidentally, ICO is identical to ISO in the limit of $\mu \to 0$. The ICO-rule has proven to be very useful in difficult learning tasks (Porr and Wörgötter 2006). In fact this rule reliably works even with very high learning rates and will always converge if one manages to bring x_0 down to zero. One should, however, notice that ICO-learning is a form of non-Hebbian (heterosynaptic) plasticity, which may be less realistic from a biological point of view.

The second question concerns the problem to design a TD-learner that will also produce some motor output when learning starts (hence, when ω_1 is still zero).

The rephrased TD-rule in Fig. 2d will indeed do this. The output will be driven by x_0 early during learning and this could be used to induce behavior. Remember, this rule has two ways to enforce stability: Either $\langle\delta\rangle = 0$ or $x_0 = 0$, which makes this rule a strange chimera between TD and ISO, also rendering it sensitive to the numerical auto-correlation problems. The question, thus, is can we find a rule that needs to enforce $\langle\delta\rangle = 0$ only and still can produce some output at the start of learning? Figure 4b shows one novel possible such architecture called Acting-TD. Here we feed the δ-error into the x_0 input line. Note, this system achieves stability *only* for $\langle\delta\rangle = 0$ and will always produce some output v.

The learning rule reads as follows:

$$\textbf{TD-a}: \quad \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} = \mu \left(\alpha\, \overline{r}(t) + \beta\, v(t)'\right) u_1(t) \tag{22}$$

where v is recursively defined as

$$v(t) = u_1(t)\, \omega_1(t) + (\alpha\, \overline{r}(t) + \beta\, v(t)')\, \omega_0 \tag{23}$$

Therefore the overall weight change can only be written as an integro-differential equation (not shown), which cannot be solved.

Figure 5c–f show how this system behaves in the open-loop condition. As expected Acting-TD is stable for $\langle\delta\rangle = 0$ (panel c). Pure Hebbian learning is implemented by this rule (panel e), even when attenuating the reward input with factor α. The feedback of δ into the input of the neuron, however, leads to a potential destabilization. To show this we have introduced the amplification factor β into the derivative of the output v'. If this pathway enters the neuron with β larger than one ($\beta > 1$), destabilization occurs as shown in Figs. 5d. For values of $\beta \leq 1$, the circuit, however, remains stable. Figure. 5f shows what happens when we invert the x_0 pathway, by setting its synaptic weight ω_0 to negative values. This leads to an unstable situation during weight growth and to a strong drop as soon as $\langle\delta\rangle = 0$.

Does this circuit produce reasonable behavioral output? At least we can make some statements about the shape of the output at the start and the end of learning. At the start we have $\omega_1 = 0$ and we receive $v = \omega_0 \overline{r}$, which makes sense. When the system has converged ($\langle\delta\rangle = 0$) we get $v = \omega_1 u_1$, which essentially amounts to a process of stimulus substitution as required from such models. The $\langle\delta\rangle = 0$ condition can also be written (though disregarding the averaging $\langle\rangle$) as $0 = \overline{r} - \omega_1 u_1'$. Hence we find that, if converged, $r = -\omega_1 u_1'$ or, more specifically, $\langle r\rangle_{t\geq t_c} = -\langle\omega_1 u_1'\rangle_{t\geq t_c}$ should be fulfilled.

6 Symmetrical learning rules

So far we were able to address the problem of divergence, which was present in the basic CL-rules (Fig. 2) having achieved stable temporal sequence learning when employing the ICO-rule. Next we would like to ask if it is possible

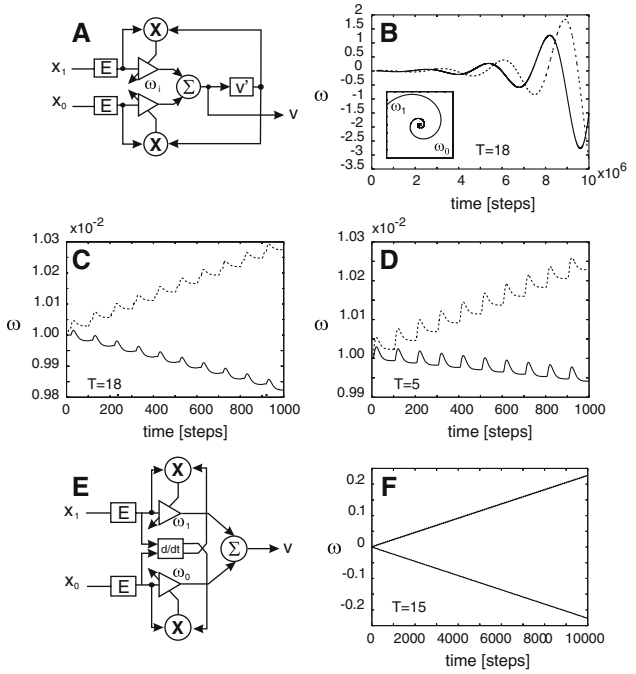

Fig. 6 Symmetrical architectures of ISO **a**–**d** and ICO **e**, **f** where both weights ω_1 and ω_0 change. Learning rate was $\mu = 10^{-4}$. Bandpass parameters were $a = 0.006$, $b = 0.0066$ and $\sigma = 0.006$. Panel **b** shows the oscillating development of both weights in ISO learning starting with $\omega_0 = \omega_1 = 0.01$ and a temporal difference of $T = 18$. The spiral shaped inset indicates that phase is constant. Panel **c** ($T = 18$) and **d** ($T = 5$) depict how the choice of T influences symmetry. In contrast to ISO learning, weights in ICO Learning **e** are symmetrical independent of the parameters used **f**

to implement learning (LTP) at one synapse and unlearning (LTD) at the other synapse *at the same time*. In principle this should be possible because one synapse experiences $+T$ while the other experiences $-T$ for any given input pair. Thus, causality is inverted for the two synapses and with the right design the one should grow, while the other would shrink.

Clearly TD-learning is designed in an asymmetrical way and cannot easily be symmetrized. This is different for the CL rules.

Figure. 6a and e show two isotropic setups. The one (Fig. 6a) for ISO-learning (which gave this rule its name) and the other for ICO-learning.

The learning rule for the coupled ISO-learning case is

$$\textbf{ISO-sym}: \quad \begin{aligned} \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} &= u_1(t)\, u_1(t)'\, \omega_1(t) \\ &\quad + u_1(t)\, u_0(t)'\, \omega_0(t) \\ \frac{\mathrm{d}\omega_0(t)}{\mathrm{d}t} &= u_0(t)\, u_1(t)'\, \omega_1(t) \\ &\quad + u_0(t)\, u_0(t)'\, \omega_0(t) \end{aligned} \tag{24}$$

Unfortunately, this system of differential equations cannot be solved analytically.

In the case of ICO-learning the rules write as follows:

$$\textbf{ICO-sym}: \quad \begin{aligned} \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} &= u_1(t)\, u_0(t)'\, \omega_0(t) \\ \frac{\mathrm{d}\omega_0(t)}{\mathrm{d}t} &= u_0(t)\, u_1(t)'\, \omega_1(t) \end{aligned} \tag{25}$$

Here we can solve the weight change analytically to

$$\textbf{ICO-sym}: \quad \begin{aligned} \Delta\omega_1 &= \tfrac{b-a}{a+b}\, \tfrac{\mathrm{sign}(T)}{2\sigma^2}\, h(|T|) \\ \Delta\omega_0 &= -\tfrac{b-a}{a+b}\, \tfrac{\mathrm{sign}(T)}{2\sigma^2}\, h(|T|) \end{aligned} \tag{26}$$

For ISO-learning, naively one would expect that with the same starting weights ($\omega_0 = \omega_1$) one should get exactly anti-symmetrical learning, because the positive influence of $+T$ at one synapse should find its exact counterpart at the other synapse which experiences $-T$. At least the learning curve in Fig. 3c, seems to suggest this. However, this is *in general not* the case due to the fact that the overlap of the filter function is not symmetrical relative to the pulse pair. Hence, one does indeed find that one synapse grows while the other shrinks, but *not* in an anti-symmetrical way. This is shown in Fig. 6b, where we ran the analysis for very many time-steps. Both weights behave anti-cyclically and the observed oscillation grows as a consequence of the increasingly more influential auto-correlation term. Phase relationship between both synapses, however, remains the same (see inset in Fig. 6b). The result in Fig. 6b was obtained with an optimal choice of parameters ω_0, ω_1 and T leading to an almost ideal anti-symmetrical learning, shown in Fig. 6c, which depicts the first learning steps. As mentioned, this situation is not generic. More often much more asymmetrical situations like in Fig. 6d are observed, which was obtained with the same setup only using a different value for T.

Symmetrical ICO-learning (Fig. 6e) produces a linear phase-relation, which is not shown here, but Fig. 6f shows instead that both weights develop in an accurate anti-symmetrical way. Weight development will exactly follow the weight change curve in Fig. 5b. There is however a problem. Symmetrical ICO-learning does not anymore have one shared control parameter for the weight change, which for symmetrical ISO-learning was the derivative of the output. For symmetrical ICO-learning, two totally independent control parameters exist (the derivatives of the inputs). This can possibly lead to problems when wanting to control behavior with such a symmetrized ICO-rule.

7 Stabilizing ISO-learning with a third factor

ICO-learning is very stable but, as mentioned above, it is a form of non-Hebbian learning, where the output does not influence the learning. This may be undesirable in certain cases. Therefore, we undertook the effort to stabilize our

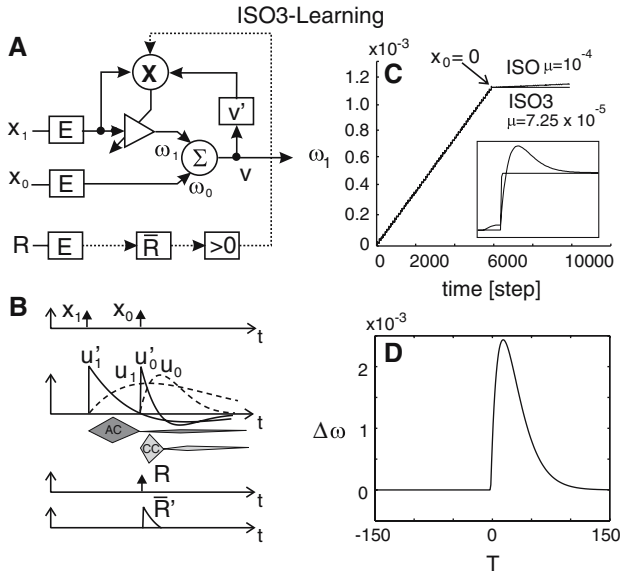

Fig. 7 Architecture and open loop simulations of ISO3-Learning. Different from ISO Learning, **a** contains a third factor R (relevance signal), which controls the moment of the weight change. **b** Shows the signal structure and gives a graphical indication of the strengths of the auto- (AC) and cross-correlation (CC) contributions. **c** compares the weight development of ISO with ISO3 for the signal structure and filter characteristics shown in **b**. ISO3 does not produce any upwards drift anymore when switching x_0 off. **d** Shows the weight change curve. The parameters of the band-bass used for R were $a = 0.6$, $b = 0.66$, $\sigma = 0.06$ and $T_R = T$ in all cases

(Hebbian) ISO-learning rule (Porr and Wörgötter 2007). This can be achieved using a third factor, which we call the "relevance signal" R (Fig. 7a). For practical purposes most of the time we set it equal to x_0, but one should realize that—like the reward line in TD-learning—R is indeed an independent signal. The signal R is meant to arise when for the animal/agent a behaviorally relevant event occurs.

The learning rule is similar to the ISO-rule (Eq. 8):

$$\textbf{ISO3}: \quad \frac{\mathrm{d}\omega_1(t)}{\mathrm{d}t} = \mu\, v'(t)\, u_1(t) \overline{R}(t)' \tag{27}$$

And therefore the weight change is

$$\textbf{ISO3}: \quad \begin{aligned} \Delta\omega_1^{cc} &= \omega_0 \int_0^\infty h(t)h'(t-T)h'_{a_R,b_R}(t-T_R)\mathrm{d}t \\ \Delta\omega_1^{ac} &= \omega_1 \int_0^\infty h(t)h'(t)h'_{a_R,b_R}(t-T_R)\mathrm{d}t \end{aligned} \tag{28}$$

where we introduced a new time interval T_R, which regulates the timing of the third factor, and a very narrow filter signal with appropriate values of a_R and b_R (here we use $a = 0.6$, $b = 0.66$) to define the third factor. It is interesting that the auto-correlation term for ISO3 is in general unequal zero when using two filters pointing to a possible instability.

Figure 7b shows the signal structure. Let us assume that u_1 reaches its maximum exactly at T. As $v'(t) = u_1'(t)$, $t < T$, we have $\lim_{t \to T_-} v(t)' = u_1'(T) = 0$. This is the situation depicted in panel (b). If we furthermore assume that the R signal is very short (e.g., using a delta-pulse for R) and that it also happens at T, then also the learning only takes place at this moment in time. Under these conditions, it is easy to see that we have totally eliminated the auto-correlation contribution. The outcome of panel c is obtained under these condition and ISO3 is stable (compare ISO and ISO3 in Fig. 7c). The inset shows again the relaxation behavior of ISO3 (step) for a single pulse-pair in comparison to ISO (curved), which demonstrates instantaneous relaxation of ISO3. Clearly this example is constructed as T is usually unknown such that $\lim_{t \to T_-} v(t)' = 0$ cannot be generally assured. Hence, it seems we have not gained anything so far by introducing ISO3. However, the situation changes when using a filter bank to spread signal x_1 out in time (see also section on filter banks below). Then one can prove that the condition $\lim_{t \to T_-} v(t)' = 0$ will self-emerge as a consequence of the learning when using enough filters (Porr and Wörgötter 2007). Thus, when using a filter bank ISO3 becomes a very stable method, indeed. For ISO3 we receive the learning curve in Fig. 7d, which now only carries an LTP component. This behavior compares to TD-learning (see Fig. 3d).

8 Results when using a filter bank

The usefulness of all these rules as presented so far remains limited as most of the time the interval T between incoming inputs is not good enough known and might, in a behaving agent, even vary to quite some degree. Hence it is required to use a set of different eligibility traces $E_1^{1,\ldots,N}$ to make sure that the earlier input is spread out over a long enough time such that the later input (x_0 or r) can be correlated to it. Figure 8a depicts such a filter bank architecture for the ISO rule and panel b shows how the signals $u_1^{1,\ldots,N}$ look like for a set of filters h.

Interestingly, convergence properties for the CL-rules are theoretically not affected when using a filter bank. It can be shown that for the ISO rule a set of filters h exist that fulfills a certain orthogonality criteria and ISO will then still converge for $x_0 = 0$ (Porr and Wörgötter 2003; Porr et al. 2003). The problem is that this is only an existence proof and nothing is currently known of how to actually construct this filter bank. Hence, when wanting to use ISO one has to fall back onto heuristic assumptions for the filter bank. Generally, however, this leads to the situation that the error-sensitivity of ISO can become larger, rendering this rule instable. The conclusion is that, while it has been the first differential Hebbian learning rule to be used in closed loop behavioral control

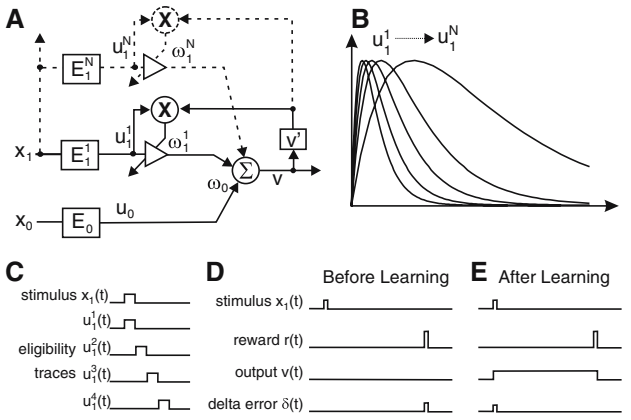

Fig. 8 Filter bank approaches for CL and RL. **a** Schematic diagram of a filter bank implementation of ISO-learning. **b** Signal structure for signals u. **c** Serial compound representation: the simulated stimulus $x(t)$ is split into traces $u_1^{1\cdots n}(t)$ with only one non-zero value in time, which cover the duration until the reward is presented. Each of these traces has its own weight $\omega_1^{1\cdots n}$ which gets updated after each time step according to the delta error δ. **d** Using TD-learning the output is zero before learning, and thus the delta error δ peaks at the same time step the reward occurs. **e** After learning δ has moved forward in front of the reward, while the output v arises with the stimulus and ends with the reward (Suri 2002; Wörgötter and Porr 2005)

(Porr and Wörgötter 2003), ISO should only be used with great care anymore.

The properties of ICO and ISO3 are better. Both rules are stable and stability for $x_0 = 0$ can be mathematically proved for both rules even when using filter banks (Porr and Wörgötter 2006, 2007). These rules have now been successfully tested in a variety of different applications (Porr and Wörgötter 2006; Kolodziejski et al. 2006, 2007; Manoonpong et al. 2007) and even chains of learning neurons can be constructed in a convergent way (Kulvicius et al. 2007).

Neuronal implementations of TD-learning generally solve the problem of unknown T by using a so called serial compound representation (Suri and Schultz 1998, 2001; Suri et al. 2001; Suri 2002). This is depicted in Fig. 8c. Essentially it represents a tapped delay line by which the earlier stimulus is spread out across several input lines $u_1^{1,\ldots,N}$ and for which one can even graphically show that TD will converge (Fig. 8d, e). This approach has been discussed in greater detail in Wörgötter and Porr (2005). Such a rigid pulse protocol is not biologically realistic, though, and approaches exist which replace this by "smoother functions", similar to the filter bank approach discussed above (Suri and Schultz 1998). These signals do not affect the δ-error, and convergence can still be assured for $\langle\delta\rangle = 0$. As the filter response enter the learning via v', it remains unclear of how to construct more realistic filter banks for TD-learning which in a behaving system will still allow approaching $\langle\delta\rangle = 0$ in a reliable way. So far there is no theory existing for this.

9 Discussion

The main goal of this paper was to provide a road map through the different basic as well as extended RL and CL learning rules showing their fundamental mathematical properties in an open loop situation (hence without behavioral feedback). Table. 1 summarizes these observations. To this end, in the previous sections we have gone through a variety a temporal sequence learning rules first introducing their basic version and later some more advanced modifications. Three aspects were in the fore-front of the discussion: (a) convergence properties, (b) symmetry (hence, learning versus unlearning) and (c) the behavior of these rules when using a filter bank.

Concerning (a) we have found that all TD-rules are stable if one can assure $\delta = 0$ or leastwise $\langle\delta\rangle_{t\geq t_c} = 0$. From the other rules the ICO and ISO3 rules are stable for $x_0 = 0$. ISO is only theoretically stable but strongly affected by numerical artifacts.

Concerning (b) we can state that only the ICO-rule is truly symmetrical. Hence, with this rule it would be possible to implement learning at one input (the one that experiences $+T$) and unlearning at the other input (the one that experiences $-T$) *at the same time*. With ISO this is to some degree possible but growth and shrinkage are not the same, while all TD-rules are by construction not symmetrical.

Concerning (c) we observe that ICO and ISO3 will maintain their convergence properties for $x_0 = 0$. TD-learning will converge when using a serial compound representation (see Fig. 8c). This sitation is very similar to a state-space tiling performed in machine learning version of TD. On the other hand, no clear theoretical results exist for TD, when using different, more neuron-compatible filters for the eligibility traces.

An additional consideration concerns the difference between input- and output-control. The CL-rules enforce convergence via $x_0 = 0$, which represents an input condition. TD learning converges via output control ($\langle\delta\rangle_{t\geq t_c} = 0$). This difference may lead also to differences for control applications as discussed elsewhere (Wörgötter and Porr 2005).

9.1 Biological relevance

To what degree are the above discussed models related to temporal sequence learning mechanisms in the brain. At first we notice that certainly all of them are at a much higher level of abstraction as compared to the biophysics of synapses. But, for example, we note that the learning curve of ISO or ICO learning resembles curves measured for spike timing-dependent plasticity (Markram et al. 1997; Magee and Johnston 1997; Bi and Poo 2001). Hence it is possible to model STDP with such a formalism (Saudargiene et al. 2004; Roberts 1999).

Table 1 Overview over all learning rules discussed in this paper

Class		Two Inputs			
Type	Name	Output: v	Rule: $\frac{\mathrm{d}w_1}{\mathrm{d}t}$	Convergence	Comment
CL	S&B	$w_0x_0 + w_1x_1$	u_1v'	divergent	stimulus substitution
	ISO	$w_0u_0 + w_1u_1{}^*$	u_1v'	$x_0 = 0$, unstable!	symmetric, diff. Hebb
	ICO	$w_0u_0 + w_1u_1{}^*$	$u_0'u_1$	$x_0 = 0$	diff. Heterosyn.
	ISO3	$w_0u_0 + w_1u_1{}^*$	$u_1v'\overline{R}'$	$x_0 = 0$, $(R = 0)$	3-factor diff. Hebb
RL	TD	w_1x_1	$\delta_r u_1 = \bar{r}\,u_1 + u_1v'$	$\langle\delta\rangle = 0$	critic only, no actions
	TD-r	$w_0u_0 + w_1u_1{}^*$	$\delta_{u_0}u_1 = u_0u_1 + u_1v'$	$\langle\delta\rangle = 0$, $x_0 = 0$	mixed Hebb + diff. Hebb
	TD-a	$w_0\delta_r' + w_1u_1$	$\delta_r u_1 = \bar{r}\,u_1 + u_1v'$	$\langle\delta\rangle = 0$	recursive rule
Class		Filter Bank		Summary	
Type	Name	Output: v	Convergence	General Comment	Use
CL	S&B	not applicable	not applicable	only of historical relevance	−
	ISO	$w_0u_0 + \sum_i w_1^iu_1^i{}^*$	$x_0 = 0$ for certain unknown h_i	unstable! As optimal h_i are unknown, convergence cannot be guaranteed, input control	−
	ICO	$w_0u_0 + \sum_i w_1^iu_1^i{}^*$	$x_0 = 0$	robust, heterosynaptic, input control	+
	ISO3	$w_0u_0 + \sum_i w_1^iu_1^i{}^*$	$x_0 = 0$, $(R = 0)$	robust, input control	+
RL	TD	serial compound rep.	$\langle\delta\rangle = 0$	robust, output control	+
	TD-r	not tested so far	$\langle\delta\rangle = 0$, $x_0 = 0$	undesirable mix of Hebb & diff. Hebb	−
	TD-a	not tested so far	$\langle\delta\rangle = 0$	not tested so far	?

The asterisk * depicts identical equations within one column

Furthermore, it has long been discussed that TD-learning could be related to dopaminergic responses in the brain. Especially the behavior of some cells in the substantia nigra and ventral tegmental area (VTA) suggest that they represent the δ-error of TD-learning. Models, which behave in a similar way have been made by Suri and co-workers (Montague et al. 1996; Suri and Schultz 1998, 1999, 2001; Suri et al. 2001) and we direct the reader to this literature for an in-depth discussion. The problem, with these models, however, is that it is difficult to find appropriate biophysical equivalents for the implementation of the TD-rule.

Concerning the CL-rules, there are different degrees of realism. For the ICO-rule we find that it represents a special case. This is due to the fact that ICO-learning implements plain heterosynaptic plasticity and this is found only at a few specialized synapses (Humeau et al. 2003; Tsukamoto et al. 2003). Heterosynaptic plasticity is usually associated with modulatory processes and not directly with Hebbian learning. This is different for the ISO-rule, which uses the derivative of the output to control learning and therefore represents conventional (differential) Hebbian learning.

The instability of the ISO-rule, however was the reason for us to design ISO3, which is a form of (differential) Hebbian learning using a three-factor learning rule (Miller et al. 1981). Such three-factor rules have recently also been discussed in conjunction with the Dopaminergic system of the brain (Schultz 1998). Also, since it is a Hebb-rule, it is better suited to be matched to our knowledge concerning LTP and LTD. Furthermore, we found, quite unexpectedly, that for weight stabilization ISO3 can use one interesting aspect of the behavior of dopamine cells in the substantia nigra and VTA (Schultz et al. 1997): These cells appear to learn anticipating a reward, whereby the temporal occurrence of their response shifts from (first) t_{x_0} to (later) t_{x_1}. When doing this with our relevance signal, we found that learning stops and that weights become essentially stable even without setting $x_0 = 0$ (data not shown). Bringing the average TD-error $\langle\delta\rangle_{t\geq t_c} = 0$ down to zero does require the dopamine responses to take a very specific shape whereas for stabilizing weights in ISO3 it is enough to get a somewhat sharp response at x_1 while loosing the R-signal at x_0. This seems to be better in conjunction with the properties of dopaminergic responses which do not appear to fulfill high accuracy requirements.

Thus, an experimental question now arises: Do the dopamine cells in the substantia nigra and/or VTA represent the δ-error in TD-learning or do they reflect a relevance signal to be used as a third factor in the learning?

Acknowledgment The authors acknowledge the support of the European Commission, IP-Project "PACO-PLUS" (IST-FP6-IP-027657). We are grateful to T. Kulvicius for help with some figures.

References

Arleo A, Gerstner W (2000) Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. Biol Cybern 83(3):287–299

Barto A (1995) Adaptive critics and the basal ganglia. In: Houk JC, Davis JL, Beiser DG (eds) Models of information processing in the basal ganglia.. MIT Press, Cambridge pp 215–232

Bi GQ, Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166

Boykina TB (2003) Derivatives of the dirac delta function by explicit construction of sequences. Am J Phys 71(5):462–468

Dayan P (1992) The convergence of TD(λ). Mach Learn 8(3/4):341–362

Dayan P, Abbott L (2003) Theoretical Neuroscience, Computational and mathematical modeling of neural systems. MIT Press, Cambridge

Dayan P, Seynowski T (1994) TD(λ) converges with probability 1. Mach Learn 14:295–301

Florian RV (2007) Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Comput 19:1468–1502

Foster DJ, Morris RGM, Dayan P (2000) A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus 10:1–16

Gerstner W, Kempter R, Hemmen JLvan , Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383:76–78

Hull CL (1939) The problem of stimulus equivalence in behavior theory. Psychol Rev 46:9–30

Hull CL (1943) Principles of behavior. Appleton Century Crofts, New York

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

Izhikevich E (2007) Solving the distal reward problem through linkage of stdp and dopamine signaling. Cerebral Cortex 101093/cercor/bhl152

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4:237–285

Klopf AH (1972) Brain function and adaptive systems—a heterostatic theory. Technical report, Air Force Cambridge Research Laboratories Special Report No. 133, Defense Technical Information Center, Cameron Station, Alexandria, VA 22304

Klopf AH (1982) The hedonistic neuron: a theory of memory, learning, and intelligence. Hemisphere, Washington DC

Klopf AH (1986) A drive-reinforcement model of single neuron function. In: Denker JS (ed) Neural networks for computing: AIP Conference Proceedings. American Institute of Physics, New York, vol 151

Klopf AH (1988) A neuronal model of classical conditioning. Psychobiology 16(2):85–123

Kolodziejski C, Porr B, Wörgötter F (2006) Fast, flexible and adaptive motor control achieved by pairing neuronal learning with recruitment. In: Proceedings of the fifteenth annual computational neuroscience meeting CNS*2006, Edinburgh

Kolodziejski C, Porr B, Wörgötter F (2007) Anticipative adaptive muscle control: Forward modeling with self-induced disturbances and recruitment. In: Proceedings of the fifteenth annual computational neuroscience meeting CNS*2007, Toronto

Kosco B (1986) Differential Hebbian learning. In: Denker JS (eds) Neural networks for computing: AIP Conference proceedings, American Institute of Physics, New York, vol 151

Krichmar JL, Seth AK, Nitz DA, Fleischer JG, Edelman GM (2005) Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. Neuroinformatics 3:197–222

Kulvicius T, Porr B, Wörgötter F (2007) Chained learning architectures in a simple closed-loop behavioural context. Biol Cybern. doi:10.1007/s00422-007-0176-y

Magee JC, Johnston D (1997) A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. Science 275:209–213

Manoonpong P, Geng T, Kulvicius T, Porr B, Wörgötter F (2007) Adaptive, fast walking in a biped robot under neuronal control and learning. PLoS Comput Biol 3(7):e134

Markram H, Lübke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275:213–215

Miller JD, Sanghera MK, German DC (1981) Mesencephalic dopaminergic unit activity in the behaviorally conditioned rat. Life Sci 29:1255–1263

Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive hebbian learning. Nature 377:725–728

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci 16(5):1936–1947

Pfister JP, Toyoizumi T, Barber D, Gerstner W (2006) Optimal spike-timing dependent plasticity for precise action potential firing in supervised learning. Neural Comput 18:1309–1339

Porr B, Wörgötter F (2003) Isotropic sequence order learning. Neural Comput 15:831–864

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

Porr B, Wörgötter F (2007) Learning with "relevance": Using a third factor to stabilise hebbian learning. Neural Comput (in press)

Porr B, Ferber Cvon , Wörgötter F (2003) ISO-learning approximates a solution to the inverse-controller problem in an unsupervised behavioral paradigm. Neural Comput 15:865–884

Roberts P (1999) Computational consequences of temporally asymmetric learning rules: I. differential hebbian learning. J Comput Neurosci 7(3):235–46

Santiago RA, Roberts PD, Lafferriere G (2007) Spike timing dependent plasticity implements reinforcement learning. In: Proceedings of the fifteenth annual computational neuroscience meeting CNS*2007, Toronto

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 80:1–27

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

Singh SP, Sutton RS (1996) Reinforcement learning with replacing eligibility traces. Mach Learn 22:123–158

Strösslin T, Sheynikhovich D, Chavarriaga R, Gerstner W (2005) Robust self-localisation and navigation based on hippocampal place cells. Neural Netw 18(9):1125–1140

Suri RE (2002) TD models of reward predictive responses in Dopamine neurons. Neural Netw 15(4-6):523–533

Suri RE, Schultz W (1998) Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp Brain Res 121:350–354

Suri RE, Schultz W (1999) A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neurosci 91(3):871–890

Suri RE, Schultz W (2001) Temporal difference model reproduces anticipatory neural activity. Neural Comp 13(4):841–62

Suri RE, Bargas J, Arbib MA (2001) Modeling functions of striatal dopamine modulation in learning and planning. Neurosci 103(1):65–85

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

Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44

Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. In: Gabriel M, Moore J (eds) Learning and computational neuroscience: foundation of adaptive networks, MIT Press, Cambridge

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, 2002nd edn. Bradford Books, MIT Press, Cambridge

Tsukamoto M, Yasui T, Yamada MK, Nishiyama N, Matsuki N, Ikegaya Y (2003) Mossy fibre synaptic NMDA receptors trigger non-Hebbian long-term potentiation at entorhino-CA3 synapses in the rat. J Physiol 546(3):665–675

Watkins CJCH (1989) Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge

Watkins CJCH, Dayan P (1992) Technical note: Q-Learning. Mach Learn 8:279–292

Witten IH (1977) An adaptive optimal controller for discrete-time Markov environments. Inf Control 34:86–295

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