Isotropic Sequence Order Learning using a Novel Linear Algorithm in a Closed Loop Behavioural System

Bernd Porr and Florentin Wörgötter

Dept. of Psychology, University of Stirling, Stirling, Scotland, UK

Abstract

In this article, we present an isotropic algorithm for sequence order learning. Its central goal is to learn the causal relation between two (or more) inputs in order to react to the earliest incoming signal after successful learning (like in typical classical conditioning situations). We implement this algorithm in a behaving system (a robot) thereby creating a closed loop situation where the learner's actions influence its own sensor inputs to the end of creating an autonomous agent. Autonomous behaviour implies that learning goals are internally defined within the organism's capabilities. Standard learning models for sequence learning (e.g., TD-learning) need an externally defined reward. This, however, is in conflict with the requirement of an implicitly defined internal goal in autonomous behaviour. Therefore, in this study we present a system in which the external reward is replaced by a reflex loop. This loop explicitly includes the environment. Every reflex loop has the inherent disadvantage which is that its *re*-actions occur each time just after a reflex-eliciting sensor event and thus 'too late'. However, a reflex can serve as the internal reference for sequence order learning which has the task of eliminating this disadvantage by creating earlier anticipatory actions. In our system learning is achieved by modifying synaptic weights of a linear neuron with a correlation based learning rule which involves the derivative of the neuron's output. All input lines are entirely isotropic. The synaptic weight change curve of this rule is strongly related to the temporal Hebb learning rule which was found in spike timing experiments. We find that after learning the reflex loop is replaced in functional terms with an earlier anticipatory action (and pathway). In addition, we observed that the synaptic weights stabilise as soon as the reflex remains silent.

Key words: Classical Conditioning, Reinforcement-Learning, Temporal Hebb, Temporal Difference Learning, Sensor-Motor Loop, Feed-forward compensation

1 Introduction

The standard paradigm for temporal sequence learning in animals is classical conditioning which dates back to Pawlow's revolutionary work early last century (Pavlov, 1927). His descriptive theory was later formalised by Rescorla and Wagner (Rescorla and Wagner, 1972) and then further developed by Sutton and Barto (1982) into their model of temporal difference (TD-) learning. TD-learning allows the formulation of a sound mathematical theory and is analytically treatable to a large extent. Although the mathematical framework suggested an end point in this development Klopf (1988) rightfully pointed out that the Sutton&Barto model is a typical supervised learning scheme arguing that animals can not learn in a supervised mode because they are autonomous agents.

The difference between supervised and unsupervised temporal sequence learning, to which our ISO-learning (Isotropic Sequence Order) scheme belongs, is formally reflected by the definition of "the reward": In the TD-model the reward is (in Klopf's words) an *evaluative feedback* which means that the supervisor (the environment) gives the system a feedback whether learning was successful or not. Learning in the sense of the TD-algorithm is, therefore, strongly related to standard supervised learning schemes, especially learning with the delta-rule (Widrow and Hoff, 1960).

In all of these algorithms, the reward serves as a reference signal for learning success. Mathematically this is defined by means of an error signal between learning-achievement and learning-goal. This error signal is then used to control the learning rate. Learning stops when a (close to) zero error is obtained. For example, in temporal sequence learning, if an event is successfully predicted the error is zero and the learning has reached its externally pre-defined goal.

Klopf's main argument against such learning schemes is that organisms are autonomous and that there is no external teacher. This requires a *non-evaluative feedback* (Klopf, 1988) which operates in accordance to an intrinsically defined learning goal.

In this study, this internally defined goal will be *punishment avoidance*. Initially this goal is achieved by means of a reflex loop, and the reflex-like avoidance reaction serves as reference behaviour in our system. This reference behaviour is superseded by a learning algorithm which replaces the slow feedback loop with a fast feed-forward pathway. Learning is guided in our system by non-evaluative feedback, which occurs simply as a consequence of having placed the learner in its environment. This is due to the fact that the environment effectively connects the organism's motor outputs to its sensor inputs in

a feedback-like manner, without explicit specification of a "teacher". Furthermore we demonstrate that during learning synaptic weights will stabilise in our system leading to an equilibrium between learner and environment.
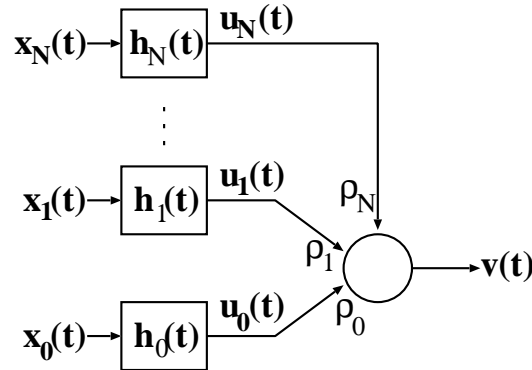
2 The neuronal circuit

Fig. 1. General form of the neural circuit. Inputs x_k are filtered by resonators with impulse response h_k and summed at v with weights ρ_k.

Fig. 1a shows the basic components of the neuronal circuit. The learning system consists of multiple inputs x_k which are first band-pass filtered by means of linear transfer functions (in the time- and Laplace-domain)

$$h(t) = \frac{1}{b} e^{at} \sin(bt) \tag{1}$$

$$H(s) = \frac{1}{(s+p)(s+p^*)} \tag{2}$$

with $a := \mathrm{Re}(p) = -\pi f/Q$ and $b := \mathrm{Im}(p) = \sqrt{(2\pi f)^2 - a^2}$, where f is the frequency of the oscillation and Q the quality. This operation is commonly observed at sensorial inputs and/or in real neuronal circuits (Shepherd, 1990). The transformed inputs u_k converge onto a single learning unit with weights ρ_k and its output is given by:

$$v = \sum_{k=0}^{N} \rho_k u_k \tag{3}$$

Note that all weights are allowed to change, which makes this setup isotropic. Since we are dealing with a behavioural paradigm this unit has the task to transform a sensor input into a motor reaction. The weight ρ_0 of the input

3

x_0 is set initially to a large value[1] and this pathway can be interpreted as the input to an unconditioned reflex loop. All other weights are associated with the conditioned stimuli. Weights change by a learning rule which uses the temporal derivative of the output:

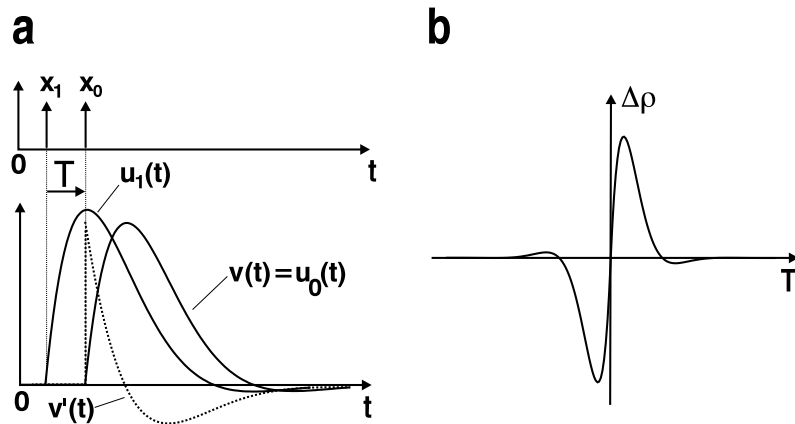$$\frac{d}{dt}\rho_j = \mu u_j v' \qquad \mu \ll 1 \tag{4}$$

Fig. 2. **a** Response curves. Curves for u show band-pass filtered responses to δ-pulse inputs ($Q = 1.0, f = 0.01$). At $t = 0$ (first pairing of pulses) we get $u_0 = v$. Note that v' has an approximately 90° phase-lead with respect to v. This arises from the band-pass filter characteristic of the system and is the basis for the predictive properties of the learning rule used. **b)** Weight change curve: same input signals as in a) but now the dependence of the weight change $\Delta\rho$ from the temporal difference T is plotted.

In order to clarify the basic idea behind this algorithm we have chosen a classical conditioning paradigm reducing the number of inputs to two: the unconditioned input x_0 with large initial weight and the conditioned stimulus x_1. Characteristic waveforms of the system in response to δ-pulse inputs are shown in Fig. 2a. Fig. 2b shows how the synaptic weight ρ_1 of the conditioned input changes assuming identical band-pass filters for two inputs which occur with a time difference of T between them. The weight change $\Delta\rho_1$ is the result of the correlation of the two resonator responses u_0 and u_1 shown in a). Weights increase for $T > 0$ (which means that a sequence of events $x_1 \to x_0$ leads to weight increase at ρ_1, whereas the reverse sequence ($x_0 \to x_1$) leads to a decrease.
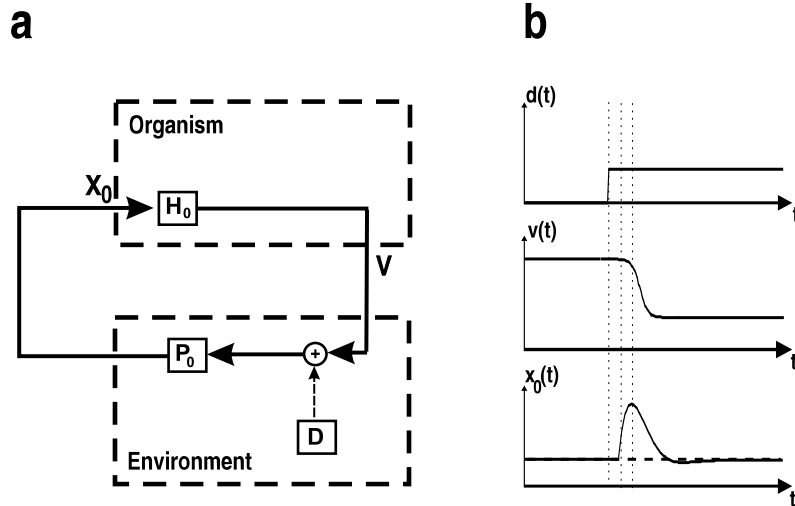
Fig. 3. **a)** Unconditioned reflex loop: the organism transfers a sensor event X_0 into a motor response V with the help of the transfer function H_0. The environment turns the motor response V again into a sensor event X_0 with the help of the transfer function P_0. In the environment there exists the disturbance D which adds its signal at $\oplus$ to the reflex loop. **b)** Signals of the reflex loop in the time domain when a disturbance $d \neq 0$ occurs. The desired state is $x_0 := 0$. The disturbance d is filtered by P_0 and appears at x_0 and is then transferred into a compensation signal at v which eliminates the disturbance at the summation point $\oplus$.

3   Behavioural feedback

Now we define a behavioural feedback reflex loop first in an abstract way and below in a real robot experiment. The generic goal of a (behavioural feedback) control-loop is to attain a desired state as good and fast as possible in this case $x_0 = 0$ (Palm, 2000). Fig. 3 shows the situation of a naive learner, who is only able to react to an unconditioned stimulus by means of a (pre-wired) reflex. In the context of control theory such a system is described by the transfer function of the system H_0, here given by the properties of sensor-motor coupling and that of the environment P_0, which is usually unknown.

Predictive learning changes this situation. The goal of predictive learning is to generate an appropriate reaction already in response to the conditioned signal which provides the disturbance D filtered by P_1 (see Fig. 4) and thereby never having to perform the reflex. When the learning goal is optimally achieved the dashed reflex pathway can be treated as functionally not anymore existent. As central observation at this stage we note that *from the perspective of the inner reflex loop* the system has now a pure feed-forward structure, because the signals do not anymore re-enter the system at the environment summation node

---

[1] In earlier experiments we have kept it constant, but this is not really necessary as we found out later.
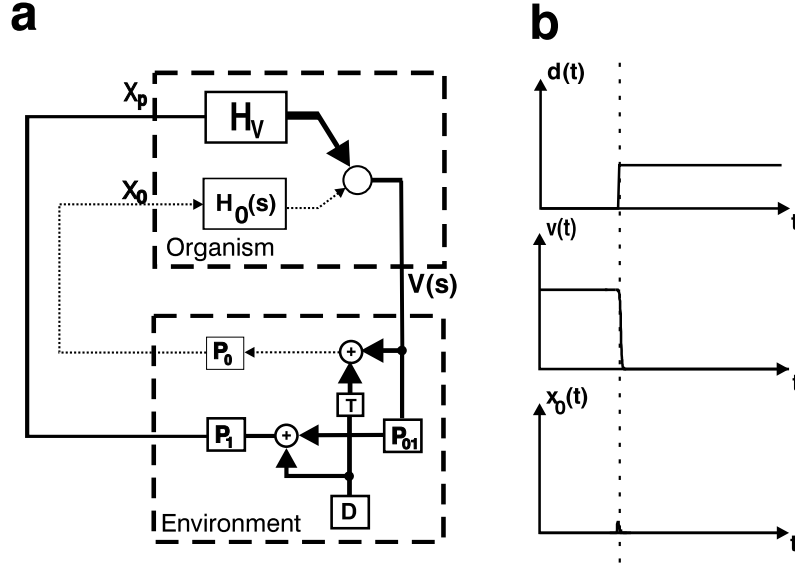
5

Fig. 4. Conditioned reflex loop. **a)** The circuit from Fig. 3 is extended by a second feedback loop via P_{01}, P_1 and X_p. The organism is now identical to the circuit in Fig. 1 where the input X_p is connected to all X_k; $k \geq 1$. Due to the delay T the disturbance D reaches the organism first via X_p and later via X_0. **b)** Time course of the signals after learning: at the moment the disturbance d has been triggered, a compensation reaction at v is elicited and the disturbance is eliminated at $\oplus$ before it can enter the input x_0.

(i.e., at the $\oplus$ before P_0). Predictive learning has effectively restructured the (slow) feedback control loop into a fast feed-forward controller. The learner's transfer function

$$H_V(s) = \sum_{k=1}^{N} \rho_k H_k \tag{5}$$

now essentially approximates the inverse of the transfer function P_1 delayed by T. At the moment the inner loop is superseded by the feed forward pathway via P_1 a new feedback loop has been generated via P_{01}. Note, however, that this loop normally only adds phase-shifts to the behaviour of the system. This will be mathematically clarified below. For now we can conclude that in most cases we can, therefore, set $P_{01} := 0$ and neglect the secondary loop.

Let us try to gain a better mathematically understanding about what underlies the situation after successful learning, i.e. when $x_0 = 0$ holds. Eq. 3 is in the Laplace domain:

$$V(s) = \rho_0 X_0 + \sum_{k=1}^{N} \rho_k X_k H_k \tag{6}$$

6

with

$$X_0(s) = P_0[V + De^{-sT}] \tag{7}$$

as the reflex pathway and

$$X_p(s) = \frac{P_1 D + P_1 P_{01} X_0 H_0}{1 - P_1 P_{01} H_V} \tag{8}$$

as the predictive pathway (see figure 4). Inserting Eq. 6 and Eq. 8 into Eq. 7 and eliminating there the $X_p(s)$ and $V(s)$ we finally get:

$$X_0(s) = e^{-sT} D + H_V \frac{P_1 D + P_1 P_{01} X_0 H_0}{1 - P_1 P_{01} H_V} \tag{9}$$

Solving for $X_0(s) = 0$ we get:

$$H_V(s) = \sum_{k=1}^{N} \rho_k H_k = -\frac{P_1^{-1} e^{-sT}}{1 - P_{01} e^{-sT}} \tag{10}$$

The numerator of this equation is the inverse transfer function of the environment together with a delay T. The denominator can be neglected because it does not provide any more poles to the transfer function $H_V(s)$, thus it can add only phase-factors. Therefore, we can (as claimed above) set $P_{01} := 0$. In general it is possible to emulate the expression on the right of Eq. 10 by a composition of appropriate transfer functions (Blinchikoff, 1976, pp.372). In our case this is achieved with a combination of band-pass filters. A universal solution, however, cannot be explicitly given because it depends on the actual shape of the environment. We were, however, able to prove that the learning rule converges if T is known and some general assumptions about the set of functions H are made (like a specific orthogonality constraint). When chosen appropriately band-pass filters obey these assumptions demonstrating that ISO-learning is convergent and that unique solutions can be found. The mathematics to prove this however, is quite intricate and cannot be included in this short article (see Porr and Wörgötter (2002) instead). For practical purposes, however, an over-complete set of band-pass filters must be used, because T is normally not constant. This thwarts the uniqueness of the solution, which is, however, of no practical consequence. Instead, the robot results shown below demonstrate that the algorithm still converges to one of the possible solutions.

Using the same mathematical framework we can also prove for two inputs ($N = 1$) that the obtained feed-forward solution leads to a stable fix-point where the weights no longer change. Without loss of generality we define the

7

transfer functions in the environment as: $P_0(s) = 1$ and $P_1(s) = 1$ to establish the simplest form of a feedback loop. As described earlier, the weight change $\Delta\rho_j$ is calculated as a correlation between the derivative of the output v and the respective input u_k, but in this case in the Laplace domain with the help of Plancherel's theorem as:

$$\Delta\rho_j = \mu\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} -i\omega V(-i\omega)U_j(i\omega)d\omega \tag{11}$$

Inserting equations 6 and 8 into 11, taking into account that after successful learning X_0 remains zero all the time and that $|Y(s)|^2 = Y(s)Y(-s)$ the weight change becomes:

$$\rho_j(T) = \frac{\mu}{2\pi}\sum_{k=1}^{N}\rho_k \cdot$$

$$\int\limits_{-\infty}^{\infty} -i\omega|X_p(i\omega)|^2 H_k(-i\omega)H_j(i\omega)d\omega \tag{12}$$

With $N = 1$ the sum consists of only one term and we get for ρ_1:

$$\rho_1(T) = \frac{\mu}{2\pi}\rho_1 \int\limits_{-\infty}^{\infty} -i\omega|X_p(i\omega)|^2|H_1(i\omega)|^2 d\omega \tag{13}$$

This integral is anti-symmetrical and thus zero (assuming that the transfer function vanish for $\omega \to \infty$). In the general case with $N > 1$ the integral becomes zero if the transfer functions are orthogonal to each other (Terrien, 1992). Eq. 12 can be used to construct these orthogonal functions. This result can, thus, be embedded into the same analytical framework for convergence & uniqueness as mentioned above, such that the assumed solution for $X_0 = 0$ would be the only one possible. In praxis, we again refer to the robot application which shows that the weights converge during learning, even with an over-complete set of H. At this point we remark that the stabilisation strictly only occurs in the case of $\mu \to 0$. We find, however, that in a practical application this still holds for values of μ which still also guarantee fast enough learning. Thus, with a reasonably low learning rate it is possible at the end to achieve a stable equilibrium between learner and environment, which can only be disturbed if new unconditioned stimuli arrive through the (anatomically still existing) dashed connections (Fig. 4).
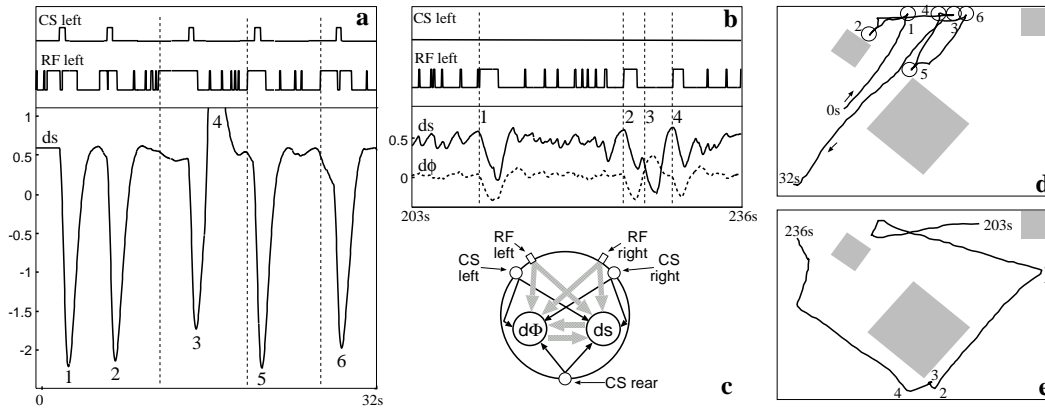
Fig. 5. The robot **c)** has 2 output neurons for speed (ds) and steering angle ($d\phi$). Normal operation is straight forward motion ($ds = const, \ d\phi = 0$). The retraction mechanism is implemented by 3 resonators ($Q = 0.6, \ f = 1$ Hz) which connect the collision sensors (CS) to the neurons ds and $d\phi$ with appropriately strong weights. Each range finder (RF) is fed into a filter bank of 10 resonators h_k with $Q = 1.0$ and $f_k = 10/k$ Hz. All outputs converge on the neurons ds and $d\phi$ with weights which are changed according the learning rule Eq. 4. A more detailed technical description goes beyond the scope of this article but is found together with a set of movies at: `http://www.cn.stir.ac.uk/predictor/real` – movie 2. **d)** Shows the robot's trajectory before learning and **a)** shows the corresponding signals of the collision sensor (CS), range finder (RF) and output of the neuron for speed (ds). The numbers correspond to the bumps in **d)**. **e)** Shows the robot's trajectory after learning and **b)** the corresponding signals.

4 Robot experiment

Fig. 5 shows results from a robot experiment. Details of the robot are described in the Appendix. Initially the robot reacts only in an unconditioned reflex-like manner with a retraction movement whenever it touches an obstacle (Fig. 5d). The nature of this behavioural paradigm, however, guarantees that the range finders of the robot will respond to an obstacle earlier than the collision sensors, which will only respond at direct contact. This temporal sequence of events is learned at two analogue summing neurons; one whose output controls the speed ds, the other the direction $d\phi$ of robot motion. The range finder signals are rather noisy (Fig. 5a,b) and, due to curvature of the motion patterns, the temporal intervals between range finder and collision-sensor signals vary within a wide range. In order to cover the large range of temporal intervals between both types of sensor signals the range finder signals were split up into ten parallel channels with different temporal transfer characteristics (Grossberg (1995), see Appendix for details). As a consequence a total of 20 learning synapses converges onto each neuron (see Fig. 5c). Despite of the poor quality of the input signals and the widely varying time intervals this system learns to perfectly correlate both sensor modalities and the robot stops colliding after about 100 seconds (compare Fig. 5d,e). The development
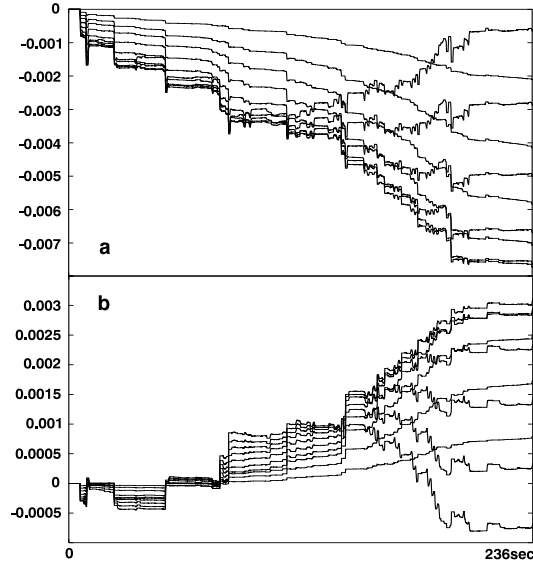
Fig. 6. Development of the weights: **a)** Shows the weights from the right range finder onto the neuron ds and **b)** onto $d\phi$.

pattern of the synaptic weights differs from trial to trial, because they develop in accordance to the sequence of sensor events, which is different for different initial situations. As a consequence different behavioural strategies are generated, for example different relations between breaking and steering. However, as expected from the theoretical considerations, we find that synaptic weights always stabilise (at different values) as soon as no more collisions occur (Fig. 6).

5   Discussion

The problem how synaptic weights stabilise during learning has puzzled biologists and theorists ever since the first correlative synaptic learning mechanisms were discovered (Hebb, 1967; Bliss and Lomo, 1973). Biochemical saturation dynamics and other mechanism have been suspected to be involved in this process (Abraham and Tate, 1997). Our results show that, during temporal sequence learning, weight stabilisation occurs in our system generically by coupling the organism with its environment.

The introduction of a novel learning rule for time-continuous signals in a poly-synaptic system with inputs from different sensor modalities was instrumental to the design of a generic feedback loop situation which is able to treat long temporal intervals between the different signals. Weight stabilisation occurs in this case as a consequence of the learner's own reactions by which it implicitly controls the sequence of its own input events. The robot example demonstrates that stabilisation of the weights is due to properties of a natural environment.

10

There is no need to stabilise the weights within the learning algorithm or within the organism. Since the algorithm is goal oriented the weights stabilise when the goal has been reached. This is an important difference from other models: in our case the prediction is *used* to reach the *internally* defined goal, but in other models the goal is to calculate a successful prediction in relation to an *externally defined reward* (Sutton and Barto, 1982; Abbott and Blum, 1996; Schultz et al., 1997).
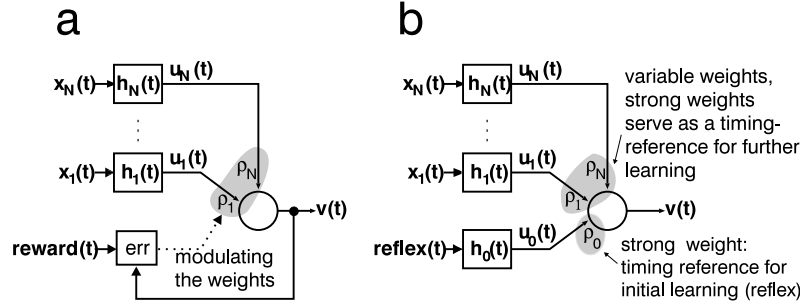

Fig. 7. Differences between a) TD-learning and b) ISO-learning.

The use of the derivative in our learning rule is related to the learning rule in TD learning (Sutton and Barto, 1982; Haruno et al., 2001) which uses the derivative as well. In the TD model the derivative is required to calculate a prediction error $(P_{t+1} - P_t)$ which is then used to change the weights. In our model the derivative of the motor output is used to utilise the predictive properties of low pass filtered signals (see the Kalman filter theory in Bozic (1994)). The second difference between the two models shows that the reward has been replaced by a feedback loop (see Fig. 7). Thus, our model needs only *one* neuronal circuit whereas the TD model needs *two*: one for the prediction and one for the error. Either models are supported by biology: for example the TD model is supported by results from Schultz et al. (1997) and the model presented here by Hauber et al. (2001). Another biological aspect is the weight change curve in our model which is reminiscent of the spike-timing dependent plasticity curves which show the change of synaptic potentials observed during long-term potentiation or depression in physiological paired-pulse experiments (Bi and Poo, 1998; Markram et al., 1997). However, further work has to be done to explore the direct biological relevance of our model on a neuronal level.

In control theory it is well known that a control loop can only exert its influence *after* the desired state has already been disturbed. In this sense reflex-reactions are slow and non-optimal. They could even lead to dangerous situations for an animal: A reflex might come too late when the initial sensor signal was generated by a life-threatening event. This represents the classical "inverse controller problem" known from engineering. One of the central goals in these disciplines is to solve this problem by replacing a (slow) feedback loop with its equivalent (faster) feed-forward controller, which emulates the inverse transfer

11

characteristic of the closed loop system (Palm, 2000). In this study we have shown that predictive learning achieves this in our system in a natural way. Related work in this field has been done by Haruno et al. (2001). They have identified the feedback feed-forward problem in motor control and using TD learning to improve arm movements in combining a feedback controller with a feed-forward controller. The difference to our model is that it is not desirable to gain autonomy in motor learning. In our model autonomy is an important aspect: the robot is allowed to develop smart (in eluding objects) but also trivial (in simply waiting in front of an object) solutions in order to solve his feedback problem.

A   Hard- and Software of the robot experiment

Hardware: A modified commercial robot ("rug warrior", 16 cm diameter) was used. Two active wheels are driven by DC motors, steering is achieved through different DC-levels. Average speed was adjusted to 0.45 m/s using a control parameter $c = 0.6$. In order to detect mechanical contact the robot has three micro-switches CS_l, CS_r, CS_b in a triangular configuration (Fig.5c). Visual signals are generated by two multiplexed, infrared emitting, active range finders RF_l, RF_r with an angle of $70°$ between them. Infrared reflection is detected by an infrared sensor centred between the emitters which operates in synchrony with them. The detection range was adjusted to $0.5 - 15.0$ cm. **Unconditioned retraction reaction:** The unconditioned retraction reaction uses only the collision sensor signals. These signals drive the output neurons in such a way that an avoidance movement with a motion vector pointing away from the site of stimulation is elicited. **Neuronal Output:** The output of the neurons is defined as: $ds = c - \rho_0^{ds}[h_r(t) * (CS_l + CS_r - CS_b)] + l_{ds}$ and $d\phi = \rho_0^{d\phi}[h_r(t) * (CS_l - CS_r)] + l_{d\phi}$. The asterisk denotes a convolution operation. The variables l_{ds} and l_{d_ϕ} represent the total sum of all learned contributions that converge onto the ds- and $d\phi$-neuron respectively. The synaptic weights in unconditioned reaction are initially set to $\rho_0^{ds} = 0.15$ and $\rho_0^{d\phi} = -0.5$.

References

Abbott, L., Blum, K., 1996. Functional significance of long-term potentiation for sequence learning and prediction. Cereb. Cortex 6, 406–416.

Abraham, W. C., Tate, W. P., 1997. Metaplasticity: A new vista across the field of synaptic plasticity. Prog.Neurobiol. 52 (4), 303–323.

Bi, G.-q., Poo, M.-m., 1998. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strengh, and postsynaptic cell type. J. Neurosci. 18 (24), 10464–10472.

Blinchikoff, H. J., 1976. Filtering in the time and frequency domain. Wiley, New York.

Bliss, T., Lomo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentrate area of the anaesthetized rabbit following stimulation of the perforant path. J Physiol 232 (2), 331–356.

Bozic, S. M., 1994. Digital and Kalman filtering: an introduction to discrete-time filtering and optimum linear estimation. E. Arnold, London.

Grossberg, S., 1995. A spectral network model of pitch perception. J Acoust Soc Am 98 (2), 862–879.

Haruno, M., Wolpert, D. M., Kawato, M., 2001. Mosaic model for sensorimotor learning and control. Neural Comp. 13, 2201–2220.

Hauber, W., Bohn, I., Grietler, C., 2001. NMDA, but not dopamin D_2 receptors in the rat nucleus accumbens are involved in guidance of the instrumental behaviour by stimuli predicting reward magnitude. J. Neurosci. 20 (16), 6282–6288.

Hebb, D. O., 1967. The organization of behavior. Science Ed., New York.

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

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

Palm, W. J., 2000. Modeling, Analysis and Control of Dynamic Systems. Wiley, New York.

Pavlov, I., 1927. Conditional Reflexes. Oxford Univ. Press, London.

Porr, B., Wörgötter, F., 2002. Learning a solution of the inverse controller problem, submitted to Physical Review Letters.

Rescorla, R., Wagner, A., 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A., Prokasy, W. (Eds.), Classical conditioning 2, current theory and research. ACC, New York, pp. 64–99.

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

Shepherd, G. M. (Ed.), 1990. The synaptic organisation of the brain. Oxford University Press, New York.

Sutton, R., Barto, A., 1982. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. Behav. Brain. Res. 4 (3), 221–235.

Terrien, C., 1992. Discrete Random Signals and Statistical Signal Processing. Prentice Hall, Englewood Cliffs, London.

Widrow, G., Hoff, M., 1960. Adaptive switching circuits. IRE WESCON Convention Record 4, 96–104.