10.1098/rsta.2003.1273

THE ROYAL SOCIETY

Isotropic-sequence-order learning in a closed-loop behavioural system

By Bernd Porr and Florentin Wörgötter

Department of Psychology, University of Stirling, Stirling FK9 4LA, UK

Published online 20 August 2003

The simplest form of sensor–motor control is obtained with a reflex. In this case the reflex can be interpreted as part of a closed-loop control paradigm which measures a sensor input and generates a motor reaction as soon as the sensor signal deviates from its desired (resting) state. This is a typical case of feedback control. However, reflex reactions are tardy, because they occur always only *after* a (for example, unpleasant) reflex-eliciting sensor event. This defines an objective problem for an organism which can only be avoided if the corresponding motor reaction is generated earlier. The goal of this study is to design a closed-loop control situation where temporal-sequence learning supersedes a tardy reflex reaction with a proactive anticipatory action. We achieve this by employing a second, earlier-occurring and causally coupled sensor event. An appropriate motor reaction to this early event prevents triggering of the original, primary reflex. Such causally coupled sensor events are common for animals, for example when smell predicts taste or when heat radiation precedes pain. We show that trying to achieve anticipatory control is a fundamentally different goal from trying to model a classical conditioning paradigm, which is an open-loop condition. To this end, we use a novel learning rule for temporal-sequence learning called isotropic-sequence-order (ISO) learning, which performs a *confounded correlation* between the primary sensor signal associated to the reflex and a predictive, earlier-occurring sensor input: this way the system learns the relation between the primary reflex and the earlier sensor input in order to create an earlier-occurring motor reaction. As a consequence of learning, the primary reflex will not be triggered any more, thereby permanently remaining in its desired resting state. In a robot application, we demonstrate that ISO learning can successfully solve the classical obstacle-avoidance task by learning to correlate a built-in reflex behaviour (retraction *after* touching) with earlier arising signals from range finders (*before* touching). Finally, we show that avoidance and attraction tasks can be combined in the same agent.

Keywords: temporal-sequence learning; control theory; reinforcement learning; inverse controller; autonomous behaviour

1. Introduction

We assume that a central goal of every autonomous agent is to maintain weak homeostasis (McFarland 1971), without which it will eventually disintegrate (or 'die'). Weak homeostasis means that the agent tries to maintain a desired input state to

One contribution of 16 to a Theme 'Biologically inspired robotics'.

the degree that it stays away from lethal boundaries (McFarland 1971). For example, an agent wants to avoid experiencing pain as far as possible. A generic way to achieve this is by reacting to a disturbance of the homeostasis with a closed-loop negative-feedback mechanism (a reflex), which will compensate for the disturbance by means of a (motor) reaction. In the example of a hot surface, the agent pulls its hand away after touching it to restore the desired state (no 'pain'). Thus, the simplest form of sensible autonomous behaviour can be obtained by designing an agent whose (re)actions are reflex based (Brooks 1989).

Such sensor–motor reflex loops represent typical feedback reaction systems, because a reflex will always be elicited only *after* a sensor event has already been encountered; as the word 'feedback' implies. The reaction delay, which is unavoidably associated with every reflex loop, can in the worst case even lead to fatal situations. Thus, in any kind of improved behaviour the acting agent will try to eliminate the disadvantage of the reflex, namely that it is always too late. For example, the experience of pain can be avoided if there is another stimulus which predicts the trigger of the pain sensor. In the example above, heat radiation and pain are causally related such that the heat-radiation signal can be used to generate an earlier motor reaction leading to a situation in which the primary reflex is no longer needed. Many other similar causal relations exist during the life of an animal; for example, between smell and taste when foraging or between vision and touch when exploring. In all of these cases, a temporal sequence of sensor events occurs which needs to be learned in order to avoid reflex reactions to the later event (see also Wolpert & Ghahramani 2000). Thus, temporal-sequence learning is a dominant aspect of animal behaviour. It requires a late event which triggers the (unwanted) reflex to which the earlier event temporally relates. The goal is to learn this specific temporal relation to issue an earlier motor reaction.

In this work, we present a new, linear and unsupervised algorithm for temporal-sequence learning, which we call isotropic-sequence-order learning (ISO learning). ISO learning solves the above problem of avoiding the late-coming reflex by replacing it with a new earlier-occurring reflex. ISO learning has the special feature that all sensor inputs are treated completely isotropically, which means that any input can drive the learning behaviour. Thus, ISO learning is completely unsupervised and the output is self-organized. Such a type of learning is highly desirable in autonomous agents which can not rely on external evaluations. Unsupervised temporal-sequence learning, however, usually leads—without additional measures taken—to rather undesirable situations for the organism, since it will probably learn arbitrary behavioural patterns. To avoid this, the ISO-learning algorithm will be placed into a non-evaluative environment (i.e. without rewards or punishments), which provides feedback from the motor output to the sensor inputs. In this behavioural context, the primary reflex prevents arbitrariness by defining an initial behavioural goal (Verschure & Voegtlin 1998).

In this study, the predefined behavioural goal is simply either an avoidance reflex only or a combination of an attraction and an avoidance reflex. These reflexes serve as the reference behaviour in our system. The reference behaviour is superseded by ISO learning, which replaces the late-feedback loop with a fast feed-forward pathway. Part of this paper is devoted to demonstrating that these qualitative observations can be embedded in a control-theoretical framework, and we will prove mathematically that the system learns a simple type of feed-forward motor control.

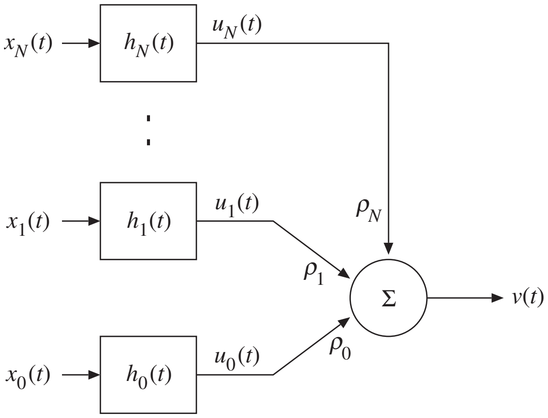

Figure 1. General form of the neural circuit in an open-loop condition. Inputs x_k are filtered by resonators with impulse response h_k and summed at v with weights ρ_k.

2. The neuronal circuit

Figure 1 shows the basic components of the neuronal circuit on which ISO learning is based. The learning system consists of multiple inputs x_k, which are first bandpass filtered by means of linear transfer functions (in the time or Laplace domain),

$$h(t) = \frac{1}{b}\mathrm{e}^{at}\sin(bt) \leftrightarrow H(s) = \frac{1}{(s+p)(s+p^*)}, \tag{2.1}$$

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 is the quality factor. 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 output given by

$$v = \sum_{k=0}^{N} \rho_k u_k, \tag{2.2}$$

with time dependence left implicit. Note that all weights are allowed to change, which makes this set-up isotropic. Since we are dealing with a behavioural paradigm, this unit has the task of transforming a sensor input into a motor reaction. Weights ρ_i change according to a learning rule which uses the temporal derivative of the output:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_j = \mu u_j v', \quad \mu \ll 1. \tag{2.3}$$

To clarify the basic idea behind this algorithm, we have chosen the simplest case for ISO learning, namely, reducing the number of inputs to two: the unconditioned input x_0 with a large initial weight and the conditioned input x_1 with an initial weight of zero. Characteristic waveforms of the system in response to δ-pulse inputs are shown in figure 2*a*. The weight change can be calculated analytically by correlating the two resonator responses of H_0 and H_1 in the Laplace space (Porr & Wörgötter 2002). Figure 2*b* shows how the synaptic weight ρ_1 of the conditioned input changes, assuming identical bandpass 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

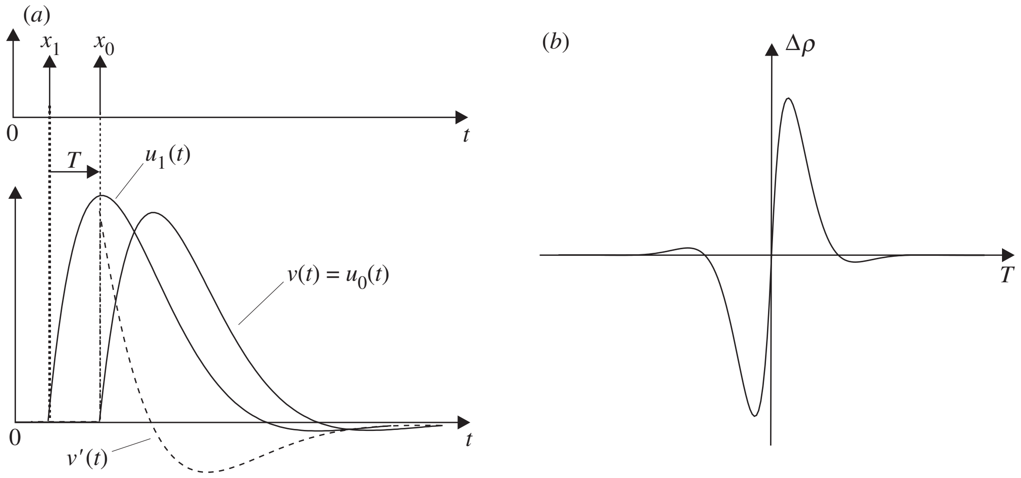

Figure 2. (*a*) Response curves. Curves for u show bandpass filtered responses to δ-pulse inputs ($Q = 1.0$, $f = 0.01$). At $t = 0$ (first pairing of pulses) we get $u_0 = v$. Note that v' has a phase lead of approximately 90° with respect to v. This arises from the bandpass filter characteristic of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

resonator responses u_0 and u_1 shown in figure 2*a*. Weights increase for $T > 0$ and decrease for $T < 0$, which means that a sequence of events $x_1 \rightarrow x_0$ leads to weight increase at ρ_1, whereas the reverse sequence $x_0 \rightarrow x_1$ leads to a decrease.

3. Behavioural feedback

We now 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 fast as possible—in this case $x_0 = 0$ (Palm 2000). Figure 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 in response to the conditioned signal which provides the disturbance D filtered by P_1 (see figure 4), thereby never having to perform the reflex. When the learning goal is optimally achieved, the dashed reflex pathway can be treated as functionally no longer existent. The learner's transfer function

$$H_V(s) = \sum_{k=1}^{N} \rho_k H_k \tag{3.1}$$

now essentially approximates the inverse of the transfer function P_1 delayed by T. At the moment that the inner loop is superseded by the feed-forward pathway via P_1, a new feedback loop is generated via P_{01}. Note, however, that this loop normally only

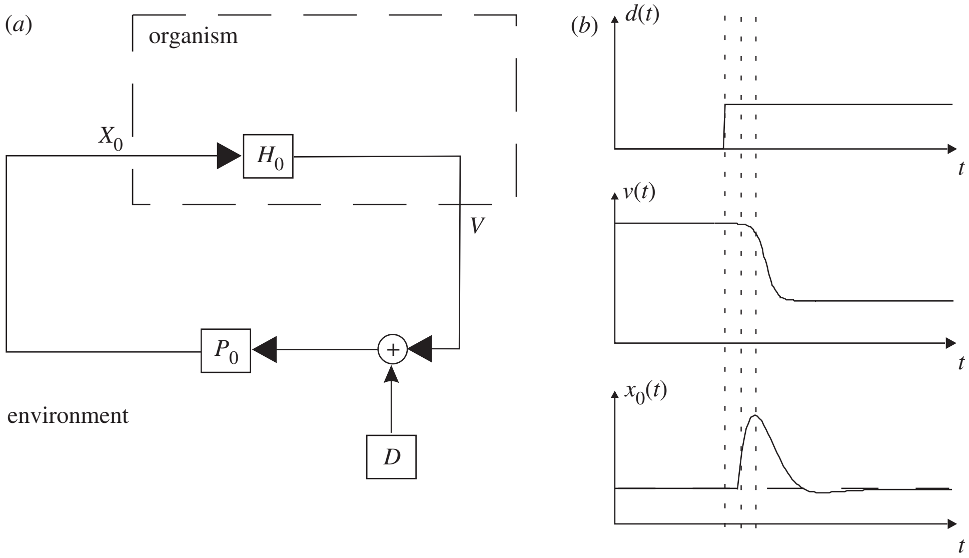

Figure 3. (*a*) Unconditioned reflex loop: the organism transfers a sensor event X_0 into a motor response V with the help of the transfer function H_0. The environment turns the motor response V into a sensor event X_0 again with the help of the transfer function P_0. In the environment there exists the disturbance D which adds its signal to the loop at $\oplus$. (*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 $\oplus$.

adds phase shifts to the behaviour of the system. This will be clarified mathematically 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 mathematical understanding about what underlies the situation after successful learning, i.e. when $x_0 = 0$ holds. In the Laplace domain, equation (2.2) becomes

$$V(s) = \rho_0 X_0 + \sum_{k=1}^{N} \rho_k X_k H_k, \tag{3.2}$$

with

$$X_0(s) = P_0[V + D\mathrm{e}^{-sT}] \tag{3.3}$$

as the reflex pathway, and

$$X_p(s) = \frac{P_1 D + P_1 P_{01} X_0 H_0}{1 - P_1 P_{01} H_V} \tag{3.4}$$

as the predictive pathway (see figure 4). Inserting equations (3.2) and (3.4) into equation (3.3) and eliminating $X_p(s)$ and $V(s)$ we finally get

$$X_0(s) = \mathrm{e}^{-sT} D + H_V \frac{P_1 D + P_1 P_{01} X_0 H_0}{1 - P_1 P_{01} H_V}. \tag{3.5}$$

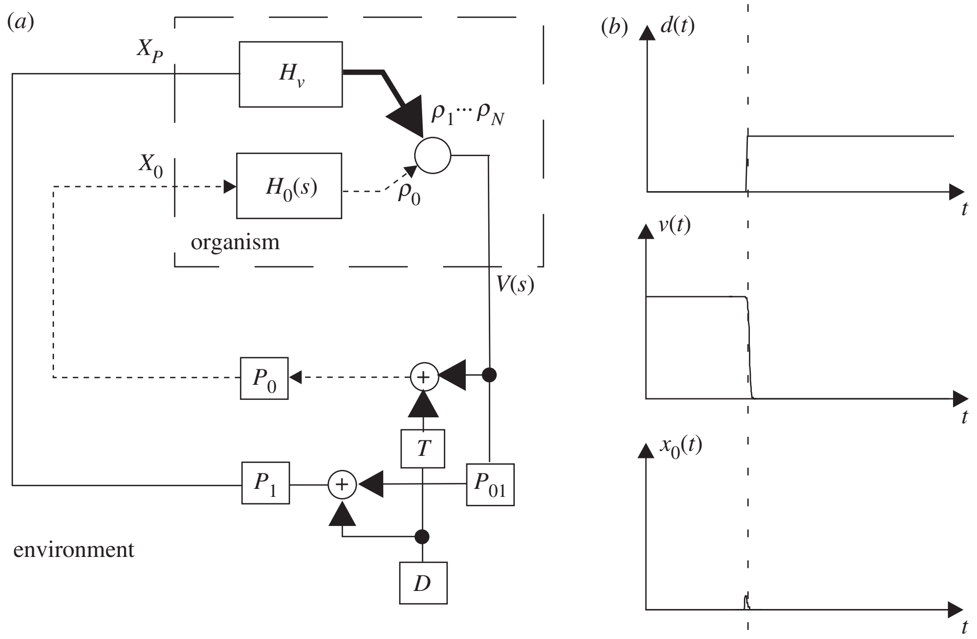

Figure 4. Conditioned reflex loop. (*a*) The circuit from figure 3 is extended by a second feedback loop via P_{01}, P_1 and X_p. The organism is now identical to the circuit in figure 1, where the input X_p is connected to all X_k, $k \geqslant 1$. Due to the delay T, the disturbance D reaches the organism 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.

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

$$H_V(s) = \sum_{k=1}^{N} \rho_k H_k = -\frac{P_1^{-1} \mathrm{e}^{-sT}}{1 - P_{01} \mathrm{e}^{-sT}}. \tag{3.6}$$

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 (Palm 2000; Stewart 1960). Therefore, we can (as claimed above) set $P_{01} := 0$. In general, it is possible to emulate the expression on the right of equation (3.6) by a composition of appropriate transfer functions (Blinchikoff 1976, p. 372).

In the following we will assume that the transfer functions are resonators. Therefore, we are looking for the weights ρ_k which lead to an approximation of equation (3.6). We start with the simplest case of $N = 1$ (one resonator) and, in addition, we assume also that the transfer function P_1 of the predictive pathway represents unfiltered throughput given by $P_1 := 1$. The right-hand side of equation (3.6) can now be developed into a Taylor series

$$-\frac{1}{\mathrm{e}^{sT}} \approx -\frac{2T^{-2}}{2T^{-2} + 2sT^{-1} + s^2} \tag{3.7}$$

and the middle part of equation (3.6) has to be explicitly written out according to equation (2.1):

$$\rho_1 H_1(s) = \frac{\rho_1}{\underbrace{pp^*}_{(2\pi f_1)^2} + s\underbrace{(p+p^*)}_{-2\pi f/Q_1} + s^2}. \tag{3.8}$$

We now can compare the coefficients of equation (3.7) with equation (3.8) and get for the parameters

$$\rho_1 = -2\frac{1}{T^2}, \qquad f_1 = \pm\frac{1}{\pi\sqrt{2}T}, \qquad Q_1 = \sqrt{\tfrac{1}{2}}. \tag{3.9}$$

These parameters can be interpreted in the following way: the maximum of the resonator with the frequency f_1 occurs approximately at time T. (It would occur exactly at T if a higher-order approximation of the delay T were used.) T is the time by which the disturbance is delayed before it enters the inner reflex loop (H_0, P_0, ρ_0, see figure 4). Thus, the response maximum of H_1 occurs at the same time as the disturbance is to be expected at the input x_1. Therefore, the signal from the resonator H_1 is suitable to eliminate the disturbance D so that x_0 is no longer triggered. To achieve an elimination of the disturbance, the right strength (and sign) of the resonator response has to be learned, which is in the end given by the weight ρ_1. We find that the value of ρ_1 renders the integral

$$\int_0^\infty \rho_1 h_1(t)\,\mathrm{d}t = -1,$$

so that it has the same energy as the disturbance D and therefore optimally counteracts it. The shape of the disturbance in the form of the δ-pulse can obviously not be achieved by a single resonator but its energy is preserved.

This solution shows that with one specific resonator it is possible to approximate one specific temporal interval T. However, if we build our system with a large enough number of resonators with different frequencies f (but constant Q), we can state that with such a system one can construct an appropriate approximation for every T in the limit of $N \to \infty$.

Now we have to show that the above approximation is stable. Let us assume that we have found a set of weights $\rho_k, k > 0$, which solves equation (3.6). The development of the weights follows,

$$\Delta\rho_j = \frac{\mu}{2\pi}\int_{-\infty}^{\infty} -\mathrm{i}\omega V(-\mathrm{i}\omega)U_j(\mathrm{i}\omega)\,\mathrm{d}\omega, \tag{3.10}$$

which is a correlation in the Laplace domain.

Now we perturb the system, substituting $\rho_j \to \rho_j + \delta\rho_j = \tilde{\rho}_j$. In order to ensure stability we must prove that the perturbation is counteracted by the weight change; thus we hope to find

$$\Delta\rho_j \sim -\delta\rho_j. \tag{3.11}$$

We need to define U and V. The first of these, U, is easy:

$$U_j = X_j H_j = \begin{cases} X_0 H_0 & \text{for } j = 0, \\ X_1 H_j & \text{for } j > 0. \end{cases} \tag{3.12}$$

However, V is more complicated. From the definition, we have

$$V = \rho_0 X_0 H_0 + X_1 \sum_{k=1}^{N} \rho_k H_k. \tag{3.13}$$

Eliminating X_0 by inserting equation (3.3), we get

$$V = \frac{\rho_0 P_0 H_0 D \mathrm{e}^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k}{1 - \rho_0 P_0 H_0}. \tag{3.14}$$

Substituting $\rho_k \to \rho_k + \delta\rho_k$ and identifying V, we get

$$\tilde{V} = V + \frac{X_1 \sum_{k=1}^{N} \delta\rho_k H_k}{1 - \rho_0 P_0 H_0}. \tag{3.15}$$

Then the weight change is found using equation (3.10),

$$\Delta\tilde{\rho}_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} -\mathrm{i}\omega \left[V^- + \frac{X_1^- \sum_{k=1}^{N} \delta\rho_k H_k^-}{1 - \rho_0 P_0^- H_0^-} \right] X_1^+ H_j^+ \,\mathrm{d}\omega, \tag{3.16}$$

where we have introduced the abbreviations '+' and '−' for the function arguments $\mathrm{i}\omega$ and $-\mathrm{i}\omega$, respectively.

We realize that the first part of this integral describes the equilibrium state condition and can be dropped; thus

$$\Delta\rho_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} \sum_{k=1}^{N} \delta\rho_k \frac{-\mathrm{i}\omega |X_1|^2 H_k^-}{1 - \rho_0 P_0^- H_0^-} H_j^+ \,\mathrm{d}\omega, \tag{3.17}$$

where for X_1 we have made use of the fact that for transfer functions in general we can write $Y^+ Y^- = |Y|^2$.

If we assume the orthogonality given by

$$0 = \int_{-\infty}^{\infty} -\mathrm{i}\omega \frac{|X_1|^2 H_j^+ H_k^-}{1 - \rho_0 P_0^- H_0^-} \,\mathrm{d}\omega \quad \text{for } k \neq j, \tag{3.18}$$

then we get

$$\Delta\rho_j = \frac{\mu}{2\pi} \delta\rho_j \int_{-\infty}^{\infty} |X_1^+|^2 |H_j^+|^2 \frac{-\mathrm{i}\omega}{1 - \rho_0 P_0^- H_0^-} \,\mathrm{d}\omega. \tag{3.19}$$

Note that this integral becomes zero if the primary reflex has been avoided ($P_0 H_0 \rho_0 = 0$). This is due to the symmetry of the transfer functions $|X_1^+|^2 |H_j^+|^2$. To prove that in all cases the integral equation (3.19) will be negative (assuring convergence), a general stability analysis will have to be performed about the inner (reflex) loop which is determined by $\rho_0 H_0 P_0$. Since we are not dealing with a specific feedback system, we choose one which is rigorously analytically tractable and which can be used as a prototype for more complex applications. To this end, we use the so-called *unity feedback loop* defined by

$$\rho_0 \in \left]-1, 0\right[, \qquad H_0 = 1, \quad P_0 = \mathrm{e}^{-s\tau}. \tag{3.20}$$

The resulting reflex loop is, thus, entirely determined by its gain ρ_0 and by the delay τ (not to be confused with T), which is the delay between the motor output

V and the sensor input X_0. The range of ρ_0 results from the demand that the reflex should be a negative-feedback loop and that it must be stable. When making these simplifications, the most important aspect of the feedback loop is still preserved, namely its delay characteristic. This property underlies the conceptual necessity for temporal-sequence learning and it is essential for any relevant mathematical treatment. The specific characteristics of some of the transfer functions, on the other hand, are secondary and can, therefore, be neglected.

Thus, equation (3.19) turns into

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} \underbrace{|DH_j|^2}_{A(\mathrm{i}\omega)} \underbrace{\frac{-\mathrm{i}\omega}{1-\rho_0 \mathrm{e}^{\mathrm{i}\omega\tau}}}_{-\mathrm{i}\omega F(-\mathrm{i}\omega)} \,\mathrm{d}\omega. \tag{3.21}$$

We now apply Plancherel's theorem (Stewart 1960) to equation (3.21) to transfer the integral back into the time domain and prove that it is negative. We have

$$\Delta\rho_j = \mu\delta\rho_j \int_0^{\infty} a(t)f'(t)\,\mathrm{d}t. \tag{3.22}$$

The function $F(s)$ of equation (3.21) is given by the transformation pair

$$F(s) = \frac{1}{1-\rho_0 \mathrm{e}^{-s\tau}}, \tag{3.23}$$

$$f(t) = (-1)^n \delta(t-n\tau), \quad n = 0, 1, 2, \dots, \tag{3.24}$$

where f represents an alternating delta function at $t = 0, \tau, 2\tau, \dots$, which starts with a positive delta pulse (Doetsch 1961). Thus, together with $-\mathrm{i}\omega$, the complete term

$$-\mathrm{i}\omega\frac{1}{1-\rho_0 \mathrm{e}^{\mathrm{i}\omega\tau}}$$

represents $f'(t)$, hence the temporal derivative of f.

The other term $A(s)$ of equation (3.21) is given by

$$A(s) = |DH_j|^2, \tag{3.25}$$

$$a(t) = \Phi[d(t) * h_j(t)], \tag{3.26}$$

where '$*$' denotes convolution and Φ is the autocorrelation function.

As a consequence of the above, we have to discuss the integral in equation (3.22) specified by equations (3.24) and (3.26). The integral should be negative to ensure stability. We know that D is short-lived with a duration shorter than τ, without which the loop system would be unstable to begin with. Thus, we can restrict the discussion of the integral to $t = 0$. We know that the autocorrelation function has a positive maximum at $t = 0$ and that the derivative f' of a δ-pulse at zero approaches $-\infty$ for $t \to 0$, $t > 0$. As a consequence, the integral is negative as required for convergence. Thus, for an orthogonal set of H_k, where we assume a unity reflex loop we have found that ISO learning will always converge if we use a disturbance that has a duration which is shorter than the loop delay τ of the unity reflex loop.

In § 4, we will use real resonator functions for H_k and H_j (which are not orthogonal) and show numerically that the system still converges. We use the unity feedback

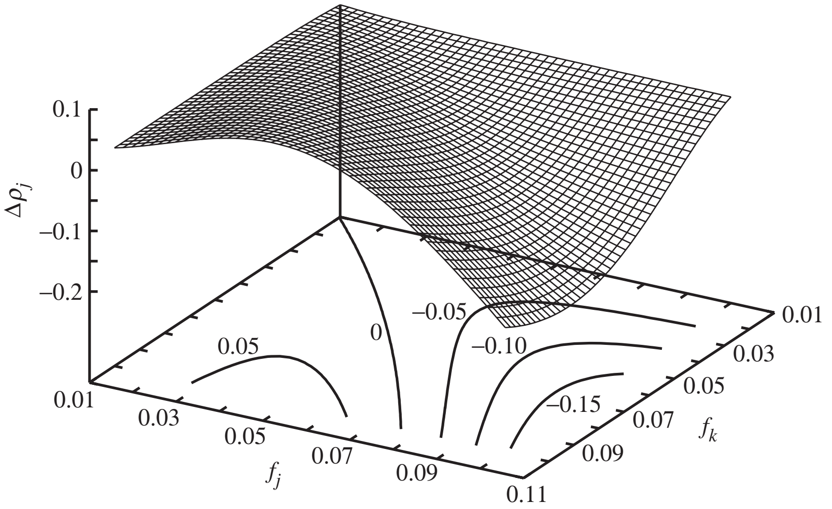

Figure 5. Numerical integration of equation (3.27). The disturbance D and the reflex loop delay τ were both set to unity. The frequencies of the resonators H_k and H_j were varied from 0.01 to 0.1 in steps of 0.001. The value of Q was set to 0.9 for both resonators. The weight of the reflex loop was $\rho_0 = -0.9$.

condition defined above (equation (3.20)) and get, for equation (3.17),

$$\Delta\rho_j = \frac{\mu}{2\pi}\sum_{k=1}^{N}\delta\rho_k\int_{-\infty}^{\infty}\frac{-\mathrm{i}\omega H_j^{+}H_k^{-}}{1-\rho_0\mathrm{e}^{\mathrm{i}\omega\tau}}\,\mathrm{d}\omega, \tag{3.27}$$

where we have set $D = 1$ which represents a δ-function as a disturbance.

Figure 5 shows the results obtained numerically for $\Delta\rho_j$ as defined in equation (3.27) in the case of a perturbation. We note that the resonators are not orthogonal, since we have for nearly all $j \neq k$ non-zero contributions. The system, however, still compensates for perturbations and thus converges for the following reason. In figure 5, we find on the diagonal that the values of the integral (equation (3.27)) are negative. This means that any perturbation at ρ_j will lead to a counterforce onto itself and, consequently to a compensation of the perturbation. However, the non-diagonal elements $k \neq j$ are non-zero. Thus, the question of stability must be rephrased into the question of how a perturbation at one given weight ρ_k will influence the other weight(s). We observe that the value of the integral (figure 5) is substantially smaller than unity everywhere. This, however, shows that any perturbation at index k will reenter the system at index j only in a strongly damped way. This process leads to a decay of any perturbation through further iterations. This strictly holds for two paired indices j and k. However, even for the complete sum in equation (3.27), which describes all cross-interference terms, we can argue that perturbations will be eliminated. This is true as long as the sum remains below unity, which is realistic given the small and sign-alternating values of the integral surface.

From this, we realize that strict orthogonality as defined in equation (3.18) is not necessary to ensure convergence. This constraint can be relaxed to the constraint that the absolute value of the sum in equation (3.27) (or equation (3.17), representing the general case) should remain below unity. Thus, for all practical purposes we can concentrate on the behaviour of the diagonal elements without having to employ an orthogonal set of H.

Now we have to consider more complex transfer functions for P_1. Up to this point, we have set $P_1 = 1$. The analytical derivation above does not show what will happen when we use more complex functions for P_1. However, when we recall that we have approximated the delay e^{-sT} with a Taylor series and matched it with the sum of resonators, we see that it is also possible to approximate more complex functions in developing them in a similar way. To allow more complex P_1, we need to make sure that P_1 can also be written as a product series of conjugate poles/zeros (hence, that it represents a standard passive transfer function). This, however, should normally be the case, because one would not expect that the environment would introduce any kind of complex signal transformation. Therefore, we can conclude that ISO learning will be able to generate anticipatory reactions in all cases where the environmental transfer function P_1 can be spelt out as a passive transfer function.

Finally, we come back to the feedback loop (H_0, P_0, ρ_0) to consider how more-complex reflex loops behave. The answer to this question was already partly given by the finding that ρ_0 has to be kept in a specific range. The restriction on ρ_0 came from the demand that the feedback loop must be stable. The same holds if we want to set up more complex feedback loops. We can expect a convergent behaviour of the complete system as soon as the reflex pathway is stable and able to maintain homeostasis. Thus, the question about the action of P_0 gets deferred to the demand of having to design an organism which has a *working* feedback loop to begin with. This means that an organism does not start as a *tabula rasa* (cf. order from noise, Der & Liebscher 2002; Haken 1995) but is already structured by reflexes (Verschure & Voegtlin 1998).

4. Robot experiment

The performance of ISO learning is demonstrated by two robot experiments. The first involves a collision-avoidance task and the second an additional attraction task. Both experiments were first simulated on a computer with a simple artificial environment. To prove the robustness of the approach, the avoidance task has also been implemented in a real robot. No substantial changes of the parameters needed to be performed when ISO learning was transferred from the simulation to the real robot.

(*a*) *Avoidance reaction by reflex avoidance*

The robot's circuit diagram is shown in figure 6. The robot has three collision sensors (CSs) and two range finders (RFs). Their signals are bandpass filtered and converge onto two neurons, which generate two different motor outputs: one controls the robot's speed ($\mathrm{d}s$) and the other its steering angle ($\mathrm{d}\phi$). The speed of the robot is set to a fixed value and its steering set to zero so that the undisturbed robot drives straight forwards. In the initial condition, before learning, the robot has only a fixed retraction behaviour (a reflex), which is triggered by the collision sensors. The range finders initially trigger no reaction (zero weights).

The signals from the range finders are able to predict the looming collision and, therefore, to generate an avoidance reaction which prevents the trigger of the collision sensors. ISO learning has the task of detecting the temporal correlation between the

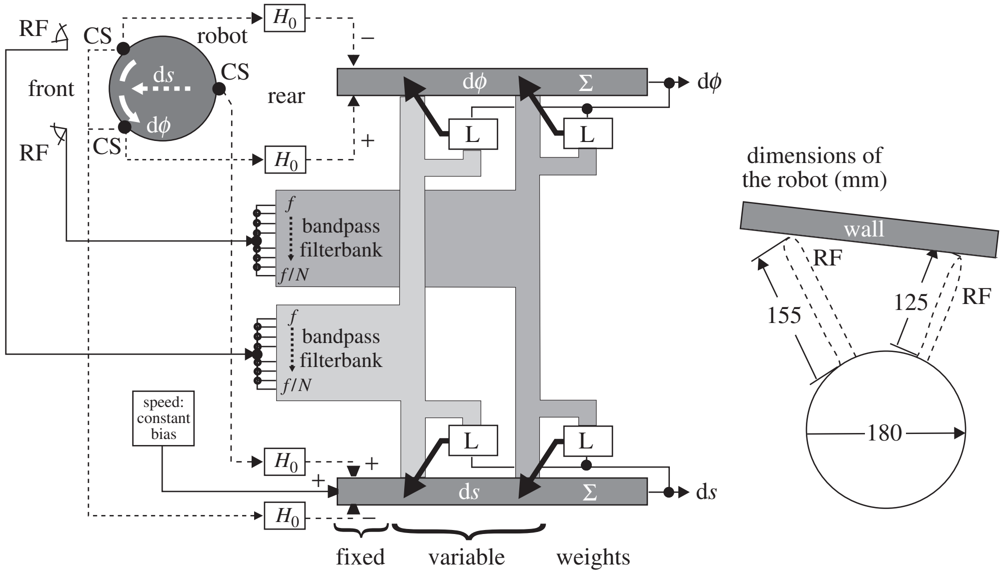

Figure 6. (*a*) Robot-circuit and (*b*) dimensions of the robot. The robot has three collision sensors (denoted ‘CS’), two range finders (‘RF’) and two output neurons (speed ds and steering angle dϕ). The output neurons are simple summation circuits ($\sum$). The reflex behaviour is triggered by the collision sensors (dotted lines) and performed by four bandpass filters (H_0 with $f_0 = 1$ Hz and $Q_0 = 0.6$). The corresponding weights are adjusted in such a way that the robot performs an appropriate retraction reaction ($\rho_0^{\mathrm{d}s} = 0.15$ and $\rho_0^{\mathrm{d}\phi} = -0.5$). The two signals from the left and the right range finders are fed into two filter-banks with $N = 10$ resonators with frequencies of $f_k = 1$ Hz$/k$; $k \geqslant 1$ and $Q = 1$ throughout. The 20 signals from the two filter banks converge on both the speed neuron (ds) and on the neuron responsible for the steering angle (dϕ). ‘L’ depicts the implementation of the learning rule (equation (2.3)) with $\mu = 0.000\,02$. Schematic (*b*) shows the dimensions of the robot and the operating ranges of the range finders. The wall demonstrates that in most of the situations both range finders are triggered when the robot is approaching a wall.

signals from the range finders and the collision sensors and thereby to generate the earlier avoidance reaction. However, the temporal delay between the range-finder signal and the collision signal is not known *a priori* and depends on the actual motion trajectory of the robot. To cope with such a wide range of temporal delays, filter banks are used in the signal pathways. Filter banks consist of 10 resonators covering a temporal interval between *ca.* 50 ms and 500 ms. These resonator signals converge onto both the speed and the steering neuron. Weight change is performed by the learning-rule equation (2.3).

Figure 7 shows the results from one robot experiment. Initially the robot reacts only in an unconditioned reflex-like manner with a retraction movement whenever it touches an obstacle (figure 7*c*). The range-finder signals are rather noisy (figure 7*a*, *b*) and, because of curvature of the motion patterns, the temporal intervals between range-finder and collision-sensor signals vary over a wide range. Despite the poor quality of the input signals and the widely varying time-intervals, this system learns to correlate perfectly both sensor modalities and the robot stops colliding after *ca.* 100 s (cf. figure 7*c*, *d*).

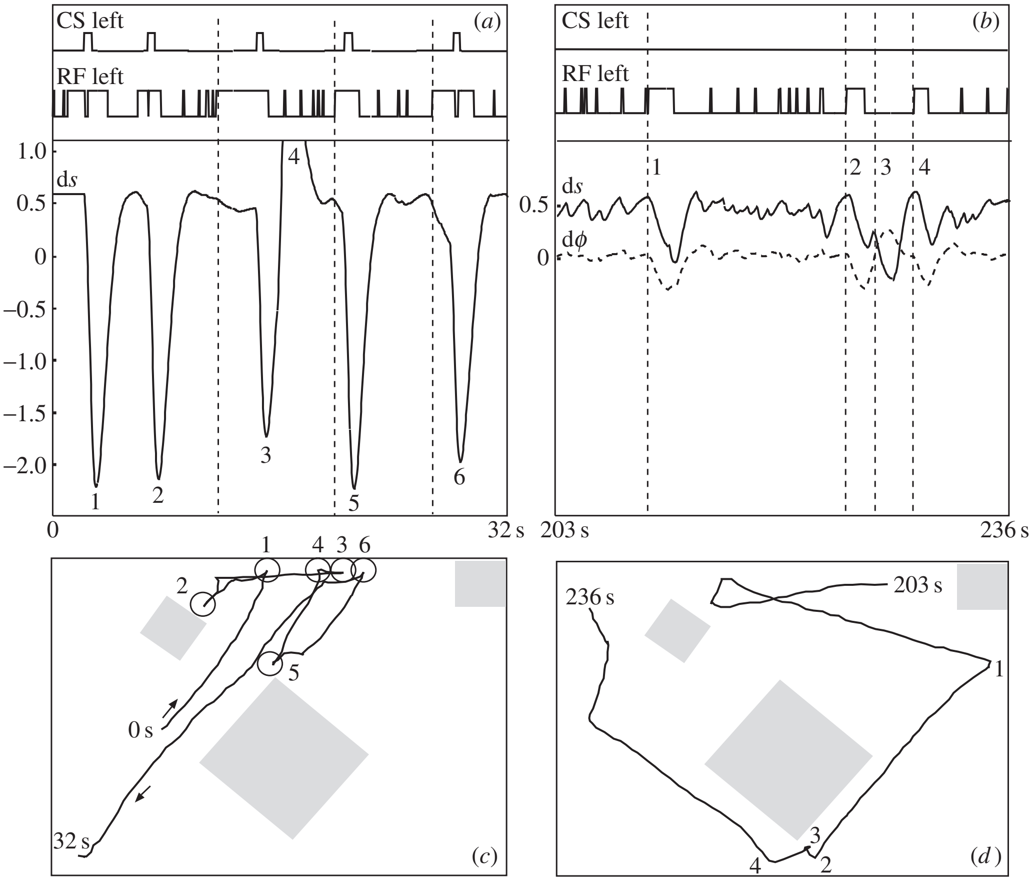

Figure 7. Trajectories (*c*) and (*d*) and the corresponding signals (*a*) and (*b*) of the robot. The films can be viewed at http://www.cn.stir.ac.uk/predictor/real, movie 2. Trace (*a*) shows the robot's trajectory before learning and (*b*) shows the trajectory after having successfully avoided the obstacles. The numbers in (*a*) and (*b*) label the events in (*c*) and (*d*), respectively.

The development of the synaptic weights differs from trial to trial, because they develop in accordance with the sequence of sensor events, which vary for different initial situations. As a consequence, different behavioural strategies are observed, for example, different relations between braking and steering.

As expected from the theoretical considerations, we find that the synaptic weights become essentially stable after the bumps have been avoided (figure 8*a*, *b*). The small remaining changes in the weights indicate that predictive learning continues: after the bump (x_0) has been avoided, learning still takes place between different range-finder inputs ($x_1, \ldots, x_N$). Learning continues until the earliest range-finder input is detected and it is able to control solely the avoidance reaction.

Figure 8*c* shows a statistical analysis of the average number of bumps which occur during a 30 000-step simulated robot run. Each curve represents the average of 400 runs with different initial conditions. The three different curves were obtained with 0, 10 or 20 obstacles, placed randomly for each individual run. Bumps were accumulated over time-intervals of 500 steps each. The 100% value represents the

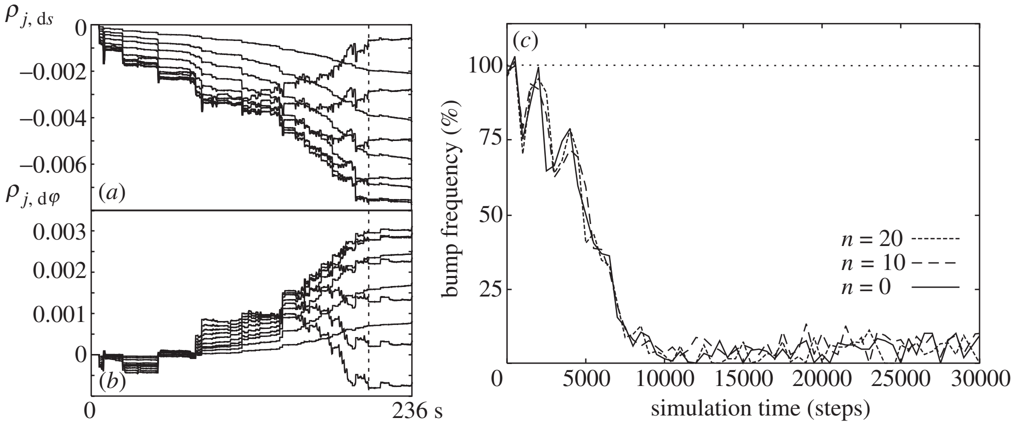

Figure 8. Development of the weights (real robot) and behavioural statistics (simulation). Part (*a*) shows the weights from the right range finder onto the neuron $\mathrm{d}s$ and (*b*) those onto $\mathrm{d}\phi$. The dotted vertical line marks the moment when the bumps have been avoided. (*c*) Shows the convergence statistics for three types of runs with $n = 0$, 10 and 20 obstacles. Runs were repeated 400 times with differently placed robots and obstacles. The learning rate was $\mu = 0.000\,0075$ (for all other parameters see figure 6). The 100% value represents the average number of bumps obtained without learning.

average number of bumps obtained in 500 time-steps without learning ($\mu := 0$). This represents on average 0.85 bumps per 500 time-steps and single run. The diagram shows that all curves fall on top of each other. This is because there is on average no correlation between consecutive bumps. All curves reach baseline after about a quarter of the total simulation time. The baseline represents approximately one bump occurring every 20 runs in each time-interval of 500 steps. Thus, individual runs will then have reached zero bumps most of the time. The few still-occurring bumps are induced by small numerical inaccuracies in the simulation (e.g. aliasing and correlation errors) and they also result from minor weight drifts which occur as a consequence of learning between the range finders as discussed above. Even in an empty playground, this leads to a few bumps after convergence. These occasional bumps are uncorrelated across runs, which leads after averaging to the observed positive baseline shift in figure 8*c*.

(*b*) *Attraction by reflex avoidance*

The robot experiment of § 4 *a* showed only an avoidance reaction. In this subsection, we will show as a computer simulation that it is also possible to add an attraction reaction to the existing avoidance reaction. A real-world application with the robot has not been made because it would require relatively elaborate peripheral hardware to simulate removable food sources. The computer simulation presented here adds an attraction reaction to the already existing avoidance reaction. Thus, both tasks are learned in conjunction.

As in the avoidance learning discussed above, the design of the reflex reaction is the crucial task as it is in the attraction-reaction presented here. While in avoid-

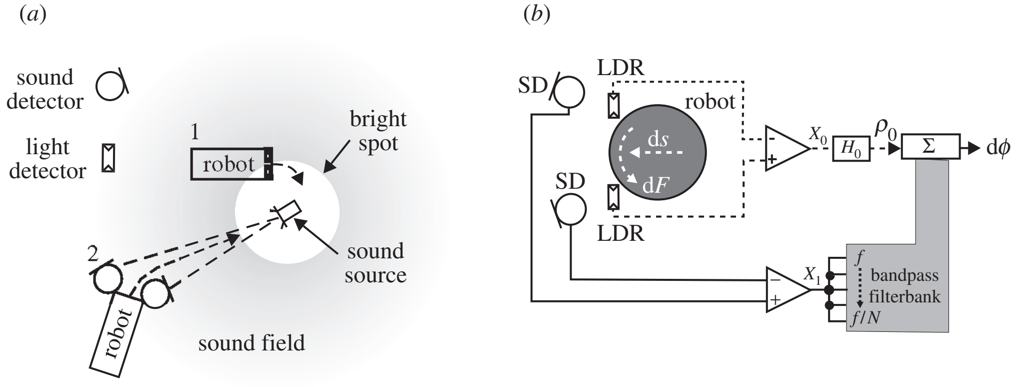

Figure 9. (*a*) Illustration of the attraction reflex and the learned behaviour. The robot is additionally equipped with two light-dependent resistors (LDRs) for light coming from above, and with two sound detectors (SDs). When one of the LDRs enters the disc, the robot turns to its centre, which causes the disc to vanish. (*b*) The two SDs enable the robot to locate an object from a distance. Additional circuitry for the computer simulation of the attraction case. The attraction task only adds a circuit for dϕ. The light detectors (LDRs, signal-range: $[0, \dots, 1]$) establish together with the resonator $H_0(f = 0.01, Q = 0.51)$ the *attraction* reflex. The weight for the reflex is set to $\rho_0 = 0.005$. The two SDs provide a signal which is inversely proportional to the distance to a sound source. The difference of the signals from both SDs is fed into a filter-bank with $f_i = 0.1/i$, $i \in [1, \dots, 5]$ and $Q = 0.51$. The learning rate was set to $\mu = 0.0002$.

ance learning the reflex represents the avoidance of an object, in attraction learning the reflex is simply the attraction behaviour towards an object. Specifically in this work, the reflex has the task of driving the robot towards the centre of a constantly illuminated area (see figure 9*a*). At the moment the robot enters the centre, the illumination vanishes and a new illuminated area appears somewhere else. Otherwise the robot might stay there forever. This process could be interpreted as targeting and eating of food.

To establish an additional reflex, the robot has been equipped with two light-dependent resistors (LDRs). The difference of these two signals drives a resonator which establishes a fixed turning reaction (see figure 9) towards the activated LDR. If both LDRs are activated identically, there will be no turning reaction (as the difference is zero).

The predictive signal for the robot is provided by a sound signal which is emitted by the centre of the illuminated area. The sound signal is detected by two sound detectors (SDs) attached to the robot. The difference of the microphone signals provides an azimuthal information for the robot about the origin of the sound source. This azimuthal information is already available from a distance and allows the robot to predict the final turning reflex. Therefore, predictive learning takes place between the sound signals and the turning reflex. Learning stops as soon as the final turning reflex is no longer triggered. This is the case when both LDRs are stimulated simultaneously, which means that the robot is heading straight for the centre of the illuminated area.

Figure 10a, b shows the trajectories before and after learning, respectively. Before learning, the robot hits the illuminated areas (marked by black discs) by chance. Entering an illuminated area (figure 10*a*) causes a reflex-like reaction and the robot

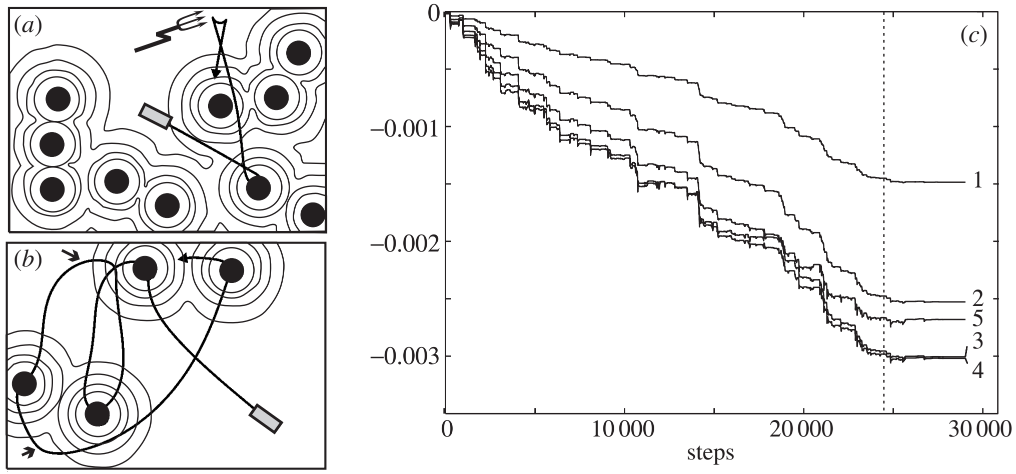

Figure 10. Trajectories of the attraction task together with the development of the weights before ($t = 0$) and after learning ($t = 21\,000, \ldots, 24\,000$). Circular fields indicate the sound field, black discs the illuminated areas marking 'food sources'. (*a*) Before learning the robot randomly finds black discs and bumps into walls (e.g. devil's arrow). Both, the black discs (1) and the walls (2) cause reflex reactions. (*b*) After learning the reflex reactions have been replaced by an avoidance reaction on the one hand (plain arrows) and by an attraction reaction on the other hand. Note that the robot's trajectory now aims directly at the centre of the black discs. Therefore, the robot enters the discs in such a way that both LDRs are triggered at the same time (reflex is no longer triggered).

does a sharp turn into the area. After such a turn, the area vanishes and a new one appears at another position. At the location marked with the devil's arrow, the robot shows the reflex reaction after a bump. From time-step 24 500 onwards, no more illuminated areas are created, so that their number decreases. After step 29 000, the playground is empty. After learning (figure 10*b*), it can be seen that the robot is directly targeting the illuminated areas and that it now hits the areas centrally. This leads to the effect that no reflex reaction ($\mathrm{d}\phi$) is caused when the robot enters the illuminated areas. Note that $\mathrm{d}\phi$ itself can be non-zero. However, in all cases $\mathrm{d}\phi$ remains constant and therefore the derivative of $\mathrm{d}\phi$ is zero. A zero derivative means that there is no learning and the weights stabilize. This can be seen in figure 10*c*, where the weights stabilize after approximately step 25 000. At that point the playground is in the condition shown in figure 10*b*, where the robot enters the illuminated areas centrally. That the weights turn out to be negative is due to the set-up of the LDRs and the SDs. For example, when the left LDR is triggered (which leads to $\mathrm{d}\phi > 0$) the input to the filter bank is negative (the left microphone is closer to the sound source than the right one). The plain arrows in figure 10*b* mark locations where the robot has learned to avoid walls.

Thus, it is also possible to construct an attraction behaviour by ISO learning. As in the avoidance case, the initial reflex defines the attraction reaction. Learning the predictive attraction behaviour again leads to the situation that the initial reflex reaction will be 'avoided' (in spite of the fact that we are in this case dealing with an attraction behaviour).

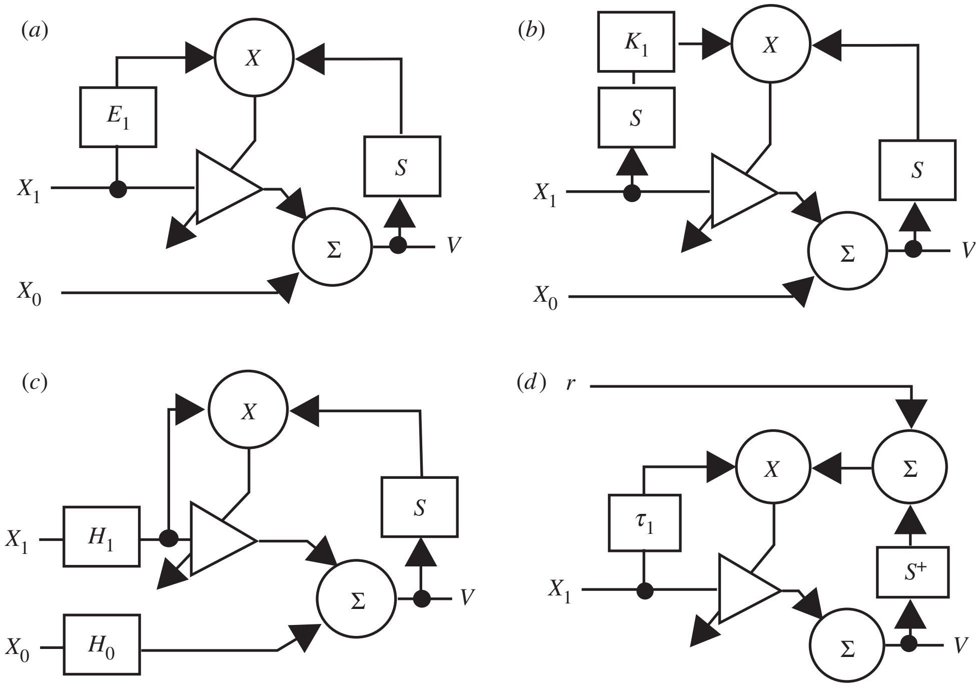

Figure 11. Differences between different models of differential Hebbian learning for $N = 1$. (*a*) Drive-reinforcement model by Sutton & Barto (1981); (*d*) temporal difference (TD) learning by Sutton (1988). The input filter E_1 is a first-order low-pass filter (eligibility trace) and τ_1 represents a delay-line. Part (*b*) shows further development of the Sutton & Barto model in (*a*), by Klopf (1986). In this model, K_1 is a tapped delay line where the coefficients are fitted to behavioural data. (*c*) ISO learning as in figure 1 with $N = 1$ together with the learning circuit to make it comparable with the other circuits. s^+ is an acausal derivative which calculates the difference between a value in the future and a value in the present.

5. Discussion

The introduction of a novel learning rule for continuous-time signals in a polysynaptic system with inputs from different sensor modalities was instrumental to the design of a generic feedback-loop arrangement which is able to treat long temporal intervals between the different signals. Weight stabilization 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 stabilization of the weights is due to the sensor–motor feedback in the environment. In other words, the weights stabilize if a specific input condition has been reached. Such a condition only makes sense in a closed-loop set-up when the motor reactions feed back to the sensor inputs (von Glasersfeld 1996, p. 151). This is an important difference from other models, which usually stabilize their weights when a specific output condition has been reached. Therefore, such models are designed for open-loop situations, in contrast to ISO learning, which is designed for situations with feedback.

The most popular models of derivative-based temporal-sequence learning ('differential Hebb', Roberts 1999) are those by Sutton (1988) and Sutton & Barto (1981, 1982). Figure 11 shows their models and our ISO learning. All models strengthen the weight ρ_1 if x_1 precedes x_0. All models use filters at the inputs. However, our model uses filtered signals for both, the learning circuit and the output, since the filtered signals are also responsible for an appropriate behaviour of the organism. In the Sutton & Barto models, the filtering affects only the learning circuitry. The central difference between these models and our approach is that the older studies aim to model experiments of classical conditioning which is an open-loop paradigm (Sutton & Barto 1987, 1990), whereas ISO learning was designed to address the closed-loop case. This is reflected in the different learning goals: in our model, the weight ρ_1 stabilizes when the input x_0 has become silent. To our knowledge, in all other temporal-sequence learning approaches, weights stabilize if the output has reached a specific condition. In the drive-reinforcement models by Sutton & Barto (1981) and Klopf (1986), this is the case if the signal at v caused by x_1 has a similar strength to that of x_0. This reflects the Rescorla–Wagner rule (Rescorla & Wagner 1972). In the case of TD learning (Sutton 1988) (or time-continuous TD (Doya 2000)), learning stops if the prediction error between reward and the output v is zero, thus if v maximally predicts r. Note that input control and output control are fundamentally different: input control can be performed strictly by the agent itself; output control requires an external observer who provides evaluations in the form of rewards and punishments. This leads to the situation that even the apparently simple problem of obstacle avoidance requires a very sophisticated treatment of the evaluative structure of the (time- and space-continuous) input space when trying to learn this task by means of traditional reinforcement control methods, like TD(λ)-control in Q-learning as discussed by Santos & Touzet (1999) and Millán *et al.* (2002).

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 is generated by a life-threatening event. Our algorithm has replaced the primary reflex by a secondary reflex which eliminates the disadvantage of the primary reflex, namely of being too late. This process could, in principle, continue (using a nested architecture) in eliminating the secondary reflex, and so on. In particular, we also observe that conflicting tasks are solved by ISO learning. In the mixed food-attraction-plus-obstacle-avoidance task above, ISO learning creates a self-adjusting equilibrium between the desire of the agent to stay away from obstacles and its simultaneous desire to approach food sources. The agent itself balances these two desires during learning and creates something like an implicit set-point which can change with changing environment or task. This leads, for example, to an agent that makes sharp turns when food sources only exist close to walls, whereas the same agent learns smooth turns when the food sources are placed in the centre of its environment. McFarland (1989) demanded that an agent should be able to resolve conflicting tasks but questioned if this could be achieved by feedback control which represents tasks explicitly by desired states (goal-directed behaviour). It seems that ISO learning offers one possible solution to this problem, as ISO learning eliminates the primary feedback loops and with them explicit representations of desired states. The secondary or higher reflex loops are an emergent property of ISO learning, allowing it to find trade-offs between conflicting primary desired states which can not be achieved at the same time. This means also that the states originally desired are no longer explicitly represented in the secondary and higher reflex loops (goal-seeking behaviour).

In addition, we have found that ISO learning generates the inverse controller of the reflex, which represents a model-free solution to the classical inverse-controller problem known from engineering. One of the central concerns 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 characteristic of the closed-loop system (Palm 2000). 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 use TD learning to improve arm movements in combining a feedback controller with a feed-forward controller. The difference between their model and our model is the point of view. In their model, the goal is defined and evaluated by an external observer. This usually requires a higher processing stage (or an experimenter/teacher) which defines the target position of the arm (desired state) and which evaluates if the target has been met or not. Therefore, in such systems there exists a unique optimal solution which has to be achieved. In our model, no higher level of processing exists which would evaluate the outcome of control. All these evaluations are performed internally in relation to the initial feedback loop. Success is achieved if the feedback loop has been superseded. This implies that there could exist more that one solution. As a result, our system develops a certain kind of basic and simple autonomy: the robot develops smart solutions (in eluding objects) but also trivial solutions (in simply waiting in front of an object) in solving its feedback problem.

The acquisition of additional useful sensorial information enables the organism to predict unwanted changes in the environment. Thus, for the organism, predicting the reflex leads to more behavioural security as compared to the situations when it had to entirely rely on its reflex *re*action. This aspect has been used by Ekdahl (2000) to define autonomous behaviour. The recruitment of more sensorial information can also be interpreted as gaining redundancy. This was pointed out by Pfeifer & Scheier (1998, 1999). Redundancy expresses itself in the context of ISO learning by the formation of nested loops. The inner reflex can always be used as a back-up if the outer (more risky) reflex fails.

Finally, we stress that learning is an active process which incorporates the sensor–motor loop as an essential part. The observed behaviour (e.g. avoiding obstacles) can be interpreted as the *recognition* of obstacles (in the context of the reflex behaviour). In our case, the robot only learns to identify obstacles. Scheier & Lambrosios (1996) used such a form of sensor–motor learning to learn *categorization* between different objects.

References

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

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

Der, R. & Liebscher, R. 2002 True autonomy from self-organised adaptivity. In *WGW'02, Proc. EPSRC/BBSRC Int. Biologically Inspired Robotics: the legacy of W. Grey Walter, HP Bristol Labs, Bristol, UK, 14–16 August 2002* (ed. R. Damper & D. Cliff).

Doetsch, G. 1961 *Guide to the applications of the Laplace and z-transforms.* London: Van Nostrand Reinhold.

Doya, K. 2000 Reinforcement learning in continuous time and space. *Neural Netw.* **12**, 219–245.

Ekdahl, B. 2000 Agents as anticipatory systems. In *Proc. 4th World Multiconference on Systemic, Cybernetics and Informatics (SCI 2000) and 6th Int. Conf. on Information Systems Analysis and Synthesis (ISAS 2000), Orlando, FL 23–26 July 2000*, pp. 133–137. Skokie, IL: IIIS.

Haken, H. 1995 *Entstehung von Biologischer Information und Ordnung.* Darmstadt: Wissenschaftliche Buchgesellschaft.

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

Klopf, A. H. 1986 A drive-reinforcement model of single neuron function. In *Neural networks for computing: AIP Conf. Proc.* (ed. J. S. Denker), vol. 151, pp. 265–270. New York: American Institute of Physics.

McFarland, D. J. 1971 *Feedback mechanisms in animal behaviour.* Academic.

McFarland, D. J. 1989 *Problems of animal behaviour.* Harlow: Longman.

Millán, J. D. R., Posentano, D. & Dedieu, E. 2002 Continuous-action *Q*-learning. *Mach. Learn.* **49**, 247–265.

Palm, W. J. 2000 *Modelling, analysis and control of dynamic systems.* Wiley.

Pfeifer, R. & Scheier, C. 1998 Representation in natural and artificial agents: a perspective from embodied cognitive science. *Z. Naturforsch.* C **53**, 480–503.

Pfeifer, R. & Scheier, C. 1999 *Understanding intelligence.* Cambridge, MA: MIT Press.

Porr, B. & Wörgötter, F. 2002 Isotropic sequence order learning using a novel linear algorithm in a closed loop behavioural system. *Biosystems* **67**, 195–202.

Rescorla, R. & Wagner, A. 1972 A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In *Classical conditioning 2. Current theory and research* (ed. A. Black & W. Prokasy), pp. 64–99. New York: Appleton-Century-Crofts.

Roberts, P. D. 1999 Temporally asymmetric learning rules. I. Differential Hebbian learning. *J. Comput. Neurosci.* **7**, 235–246.

Santos, J. M. & Touzet, C. 1999 Exploration-tuned reinforcement function. *Neurocomputing* **28**, 93–105.

Scheier, C. & Lambrosios, D. 1996 Categorization in a real-world agent using haptic exploration and active perception. In *From animals to animats, 4th Int. Conf. on Simulation of Adaptive Behaviour, Cape Cod, MA.* Cambridge, MA: MIT Press.

Shepherd, G. M. (ed.) 1990 *The synaptic organisation of the brain.* Oxford University Press.

Stewart, J. L. 1960 *Fundamentals of signal theory.* New York: McGraw-Hill.

Sutton, R. 1988 Learning to predict by method of temporal differences. *Mach. Learn.* **3**, 9–44.

Sutton, R. & Barto, A. 1981 Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* **88**, 135–170.

Sutton, R. S. & Barto, A. G. 1982 Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behav. Brain Sci.* **4**, 221–235.

Sutton, R. S. & Barto, A. G. 1987 A temporal-difference model of classical conditioning. In *Proc. 9th Ann. Conf. Cognitive Sci. Soc.*, pp. 355–378. Mahwah, NJ: Erlbaum.

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

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

von Glasersfeld, E. 1996 Learning and adaptation in constructivism. In *Critical readings on Piaget* (ed. L. Smith), pp. 22–27. London: Routledge.

Wolpert, D. M. & Ghahramani, Z. 2000 Computational principles of movement neuroscience. *Nature Neurosci. Suppl.* **3**, 1212–1217.