ISO-learning approximates a solution to the inverse-controller problem in an unsupervised behavioural paradigm

Bernd Porr*, Christian von Ferber†, and Florentin Wörgötter*

*Department of Psychology, University of Stirling, Stirling FK9 4LA, Scotland

†Theoretical Polymer Physics, Freiburg University, 79104 Freiburg, Germany

Abstract

In the previous article we have introduced an isotropic algorithm for temporal sequence learning (ISO-learning). Here we embed this algorithm into a formal non-evaluating ("teacher-free") environment which establishes a sensor-motor feedback. The system is initially guided by a fixed reflex reaction which has the objective disadvantage that it can only react *after* a disturbance has occurred. ISO-learning eliminates this disadvantage by replacing the reflex-loop reactions with earlier anticipatory actions. In this article we will analytically demonstrate that this process can be understood in terms of control theory showing that the system learns the inverse controller of its own reflex. Thereby this system is able to learn a simple form feed-forward motor control.

1 Introduction

In the previous article (Porr and Wörgötter, 2002) we have introduced a novel, linear and unsupervised algorithm for temporal sequence learning, which we called ISO-learning (isotropic sequence order learning). ISO-learning has the special feature that all sensor inputs are completely isotropic which means that any input can drive the learning behaviour. We had used the algorithm to generate robot behaviour by means of sensor inputs and motor actions. While the organism transforms sensor events into motor actions the environment passively performs the opposite and forms together with the organism a closed sensor-motor feedback loop system. Now we would like to explain the system theoretical consequences of this.

ISO-learning is completely unsupervised and the output is self-organised. Unsupervised temporal sequence learning, however, usually leads – without additional measures taken – to rather undesired situations for the organism since it can learn arbitrary behavioural patterns. A fixed reflex loop prevents arbitrariness by defining an initial behavioural goal (Verschure and Voegtlin, 1998). A reflex, however, is a typical *re*-action which will always occur only after its eliciting sensor event (Wolpert and Ghahramani, 2000). ISO-learning leads to the *functional* elimination of the reflex loop in using predictive sensorial cues and in generating appropriate anticipatory actions to prevent the triggering of the reflex. In the following, we will see that these qualitative observations can be embedded in a control theoretical framework.

In the field of control theory a reflex loop is represented by a fixed feedback loop

(McGillem and Cooper, 1984; D'Azzo, 1988; Nise, 1992; Palm, 2000). Feedback loops try to maintain a desired state by comparing the actual input value(s) with a predefined state and adjusting the output so that the desired state is optimally maintained. The main advantage of a feedback loop is that the controller only needs very *limited* knowledge about the relation between input and output (the environment). Take, for example, the typical example of a thermostat controlled central heating system. There it is necessary only to measure the temperature at the thermostat and use this to control the oven, while it is not necessary to know how much fuel needs to be burnt to get a certain temperature increase. Even this is not enough to control the heating, because the temperature increase also depends on the existing inside-outside temperature gradient and maybe on other even more elusive parameters. In general, only in idealised situations there exists sufficient prior knowledge to control a system without feedback, thus, by means of pure feed-forward control. The central advantage of such an (ideal) feed-forward controller, however, is that it acts without the feedback-induced delay. The sometimes fatally damaging sluggishness of feedback systems makes this a highly desirable feature. As a consequence, engineers try to replace feedback controllers with their equivalent feed-forward controllers wherever possible, thereby trying to solve the famous "inverse controller problem" (Nise, 1992).

In this study, we will analytically prove that ISO-learning approximates the inverse controller of a reflex when embedded in a behavioural situation where the reflex represents the reference for self-organised predictive learning.

The paper is organised in the following way. Very briefly we will summarise the main equations from the first paper, which we need here. Then, we will introduce the necessary terminology from control theory by means of discussing the reflex-loop situation. After that we will show which shape a transfer function must take in order to approximate the inverse controller of the reflex. In the next step we will demonstrate that the set of functions used in ISO-learning can indeed approximate this transfer function. Finally, we will show why the actual learning process does converge into the correct solution.

2 The ISO-learning algorithm - a brief summary

[Figure 1 about here.]

The system consists of $N+1$ linear filters h receiving inputs x and producing outputs u. The filters connect with corresponding weights ρ to one output unit v (Fig. 1). The output $v(t)$ in the time domain and its transformed equivalent $V(s)$ in the LAPLACE domain are given as:

$$v(t) = \rho_0 u_0 + \sum_{k=1}^{N} \rho_k u_k \; \leftrightarrow \; V(s) = \rho_0 U_0 + \underbrace{\sum_{k=1}^{N} \rho_k U_k}_{H_v} \qquad (1)$$

The transfer functions h shall be those of *bandpass* filters which transform a δ-pulse input into a damped oscillation. They are specified in the time- and in the LAPLACE-domain by:

$$h(t) = \frac{1}{b} e^{at} \sin(bt) \ \leftrightarrow \ H(s) = \frac{1}{(s+p)(s+p^*)} \tag{2}$$

where p^* represents the complex conjugate of the pole $p = a + ib$, with:

$$a := \mathrm{Re}(p) = -\pi f/Q, \quad b := \mathrm{Im}(p) = \sqrt{(2\pi f)^2 - a^2} \tag{3}$$

f is the frequency of the oscillation and Q the damping characteristic.

Learning takes place according to:

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

where v' is the temporal derivative of v. For a comparison of ISO-learning with other models for temporal sequence learning see Appendix B in the first paper (Porr and Wörgötter, 2002). Note that μ is very small. The integral form of this learning rule is in the LAPLACE domain given by:

$$\Delta\rho_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} -i\omega V(-i\omega) U_j(i\omega) d\omega \quad \text{with } U = XH \tag{5}$$

Note that we use indices k to denote outputs (e.g., when associated with v) while indices j denote inputs (e.g., associated with u). See (Porr and Wörgötter, 2002) for a complete description of the ISO-learning algorithm and its properties.

3 Analytical treatment of the closed loop condition

3.1 Reflex loop behaviour

[Figure 2 about here.]

Every closed loop control situation with negative feedback has a so called *desired state* and the goal of the control mechanism is to maintain (or reach) this state as good and fast as possible. In our model we assume that the desired state of the reflex feedback loop is unchanging and defined by the properties of the reflex loop. We define it as $X_0 = 0$. First we discuss the system without learning. Fig. 2a shows the situation of a learner embedded into a very simple but generic (i.e. unspecified) formal environment which has a transfer function P_0. This learner is able to react to an input only by means of a reflex.

A possible set of signals which can occur in such a system is shown in Fig. 2b. First the disturbance signal d deviates from zero, then the input x_0 senses this change $x_0 \neq 0$ and only finally the motor output v can generate a reaction in order to restore the desired state $x_0 = 0$. Thus, there is always a reaction delay in such a system.

[Figure 3 about here.]

3.2 Augmenting the reflex by temporal sequence learning

In this section we will show that the ISO-learning algorithm can approximate the inverse controller of the reflex. Fig. 3 shows how the same disturbance D elicits a sequence of sensor events: first it enters the outer loop arriving at X_1 filtered by the environment (P_1), while it arrives at X_0 only after a delay T. The goal of learning is to generate a transfer function H_v which compensates for the disturbance. The inner structure of H_v given by the ISO-learning setup is depicted by Fig. 1. The environmental transfer function P_{01} closes the outer loop.

3.2.1 General Condition

The reflex loop defines the goal of the feed-forward controller, namely that there should always be zero input at X_0. Thus, first we must show what shape the transfer function of the predictive pathway H_v (see Figs. 1 and 3) takes when we assume that $X_0 = 0$ holds. This is the necessary condition, which needs to be obeyed in order to obtain an appropriate H_v. It generally applies *regardless* of the learning algorithm used.

In the following we will omit the function argument s where possible then we can write:

$$X_0 = P_0[V + De^{-sT}] \tag{6}$$

as the reflex pathway and

$$X_1 \quad = \quad \frac{P_1 D + P_1 P_{01} X_0 H_0}{1 - P_1 P_{01} H_V} \tag{7}$$

$$H_V = \sum_{k=1}^{N} \rho_k H_k \tag{8}$$

as the predictive pathway (see figure 3). Eliminating X_1 and V we get:

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

Solving for $X_0 = 0$ leads to:

$$\begin{aligned} H_V &= \sum_{k=1}^{N} \rho_k H_k && (10) \\ &= -\frac{P_1^{-1} e^{-sT}}{1 - P_{01} e^{-sT}} && (11) \end{aligned}$$

The transfer function H_V is the overall transfer function of the predictive pathway. Eq. 10 demands that the weights ρ_k should be adjusted in such a way that Eq. 11 is obtained at the end of learning.

Eq. 11 requires interpreting. First, we consider the numerator and remember that the learning goal is to achieve $X_0 = 0$. This requires compensating the disturbance D. The disturbance, however, enters the organism only after having been filtered by the environmental transfer function P_1. Thus, compensation of D requires to undo this filtering by the term P_1^{-1}. The term P_1^{-1} is the inverse transfer function of the environment (hence "inverse controller"). The second term e^{-sT} in Eq. 11 compensates the delay between the signal in X_1 and that at X_0, when the disturbance actually enters the inner feedback loop.

Now we discuss the relevance of the denominator showing that it can be generally neglected. Transfer functions are fully described by their poles and zero-crossings.

Poles very strongly affect the behaviour of a system, while zero-crossing are phase-factors, which do not alter its general transfer characteristic (Stewart, 1960; Blinchikoff, 1976; McGillem and Cooper, 1984; Terrien, 1992; Palm, 2000). As a consequence, following methods from control theory, any transfer function may be reduced to only those terms which contain poles or zero-crossing by neglecting all other components (Sollecito and Reque, 1981; Nise, 1992).

Thus, we rewrite Eq. 11 as:

$$H_V = -P_1^{-1} e^{-sT} \frac{1}{1 - P_{01} e^{-sT}} \tag{12}$$

and analyse if the second term produces additional poles for H_V. This would happen if $1 - P_{01} e^{-sT} = 0$ holds, which is equivalent to $P_{01} = e^{sT}$. The term e^{sT}, however, is meaningless; it represents a "time-inverted delay", thus, an entity which violates causality.

As a result, there are no additional poles for H_V and in the following we are allowed to set $P_{01} := 0$ without loss of generality, thereby only neglecting possible changes in phase relationships. Thus the behaviour of H_V is apart from phase-terms entirely determined by:[1]

[1] The reader who is less familiar with control theory may find it useful to think about P_{01} also in a different way. P_{01} represents how the environmental transfer of the reaction of the system will influence the sensor X_1. Many times this influence is plainly zero from the beginning (or the connecting path can be decoupled by an appropriate

$$H_V = P_1^{-1} e^{-sT} \tag{13}$$

The last equation represents the necessary condition for the learning and we ask in the next two sections if our specific algorithm is sufficient to achieve this.

3.2.2 Solutions in the steady state case X_0=0

Here we will show by construction that already for one resonator there exists a solution which approximates Eq. 13 to the second order. Results for a forth order approximation have been numerically obtained, showing that the approximation continues to improve.

Thus, first we limit the discussion to the case of only two resonators H_0 and H_1, i.e. $N = 1$. The case with more resonators will be re-introduced at the end of this section. We will specify which parameters the resonator H_1 in the outer loop has in order to satisfy the learning goal. At first we set $P_1 = 1$, looking at the case when the environment does not alter the shape of the disturbance (but see below).

Considering Eq. 13 we have to solve:

$$-e^{-sT} = \rho_1 H_1 \tag{14}$$

The resonator H_1 has two parameters $f_1 = 1/T_1$ and Q_1 and together with its weight ρ_1 we are looking for three parameters to solve this equation.

system design). For example for a predictively acting, external (!) temperature sensor X_1 the change of the temperature of the environment due to the heating of a room is totally insignificant.

The left hand side of Eq. 14 can now be developed into a Taylor series:

$$-\frac{1}{e^{sT}} = \frac{-1}{1 + sT + \frac{1}{2}s^2T^2 + \ldots} \approx \frac{-2T^{-2}}{2T^{-2} + 2sT^{-1} + s^2} \qquad (15)$$

and the right hand side of Eq. 14 has to be explicitely written out according to Eq. 2 and 3:

$$\rho_1 H_1(s) = \frac{\rho_1}{(s+p)(s+p^*)} = \frac{\rho_1}{\underbrace{pp^*}_{(2\pi f_1)^2} + s\underbrace{(p+p^*)}_{\frac{-2\pi f}{Q_1}} + s^2} \qquad (16)$$

We can now compare the coefficients of Eq. 15 with Eq. 16 and get for the parameters:

$$\rho_1 = -\frac{2}{T^2}, \quad f_1 = \pm\frac{1}{\pi T\sqrt{2}}, \quad Q_1 = \sqrt{\frac{1}{2}} \qquad (17)$$

This result shows that for all T there exists a resonator H_1 with a weight ρ_1, which approximates e^{-sT} to the second order.

The result for f can be interpreted in the context of the previous paper (Porr and Wörgötter, 2002)). We remember that $X_0 = 0$ and hence $V = X_1 H_1$. If we consider pure δ-pulse input at X_1 (like in the simulations in Porr and Wörgötter 2002) we receive the impulse response of the resonator $h_1(t)$ at the output, thus:

$$\begin{aligned} v(t) &= \rho_1 \frac{1}{b_1} \sin(b_1 t) e^{-a_1 t} && (18) \\ &= \rho_1 T \sin(\frac{t}{T}) e^{-\frac{t}{T}} && (19) \\ &= -\frac{2}{T} \sin(\frac{t}{T}) e^{-\frac{t}{T}} && (20) \end{aligned}$$

This function has its maximum at $t_{max}^{(2)} = T\text{atan}(1)$. We can assume[2] that this is approximately equal to $t_{max}^{(2)} \approx T$. This, however, would be indicative of a response maximum which occurs exactly at the moment where the input x_0 is to be expected. We refer the reader to the previous article where this type of behaviour has indeed been observed in the simulations (Fig. 5 in Porr and Wörgötter 2002). There we have found that during learning the output has always its first maximum at the location where x_0 occurs (or would have occurred). The strength of the resonator response Eq. 19 is determined by the weight ρ_1 which is adjusted in a way that the resulting integral (Eq. 20) becomes $\int_0^\infty v(t)dt = -1$ so that it has the same energy as the δ-pulse of the disturbance D and therefore optimally counteracts it. The shape of the disturbance in form of the δ-pulse can obviously not be achieved by a single or two resonators but the energy (or the effect) is preserved.

The final stable value for ρ_1 is the main difference between the open-loop case and

[2] The relation $t_{max}^{(2)} \approx T$ could be confirmed because we performed the same Taylor approximation with $N = 2$ (leading to a forth order Taylor approximation):

$$-e^{-sT} = \frac{-1}{1 - sT + \frac{1}{2}s^2T^2 - \frac{1}{6}s^3T^3 + \frac{1}{24}s^4T^4} \tag{21}$$

$$\rho_1 H_1(s) + \rho_2 H_2(s) = \frac{\rho_1}{(s+p_1)(s+p_1^*)}\frac{\rho_2}{(s+p_2)(s+p_2^*)} \tag{22}$$

The resulting set of equations (from comparing the coefficients) has been solved numerically and we received a solution which leads to $t_{max}^{(4)} = 0.978T$. This suggests that $t_{max}^{(\infty)} = T$ is correct in the limit of $N \to \infty$.

the closed-loop case. While in the open-loop case the weight ρ_1 grows endlessly due to the lack of feedback (see the simulations in Porr and Wörgötter 2002) in the closed-loop condition the weight ρ_1 converges to a *specific value* at the moment when $x_0 = 0$ has been achieved. As a consequence the experimentally observed behaviour of the algorithms leads to a function H_v which has similar properties as that obtained from the second order Taylor approximation.

For all *practical* purposes N needs to be found in trying to resolve the tradeoff between the actually needed precision for $t_{max}^{(\infty)} \rightarrow T$ and hardware/software engineering constraints (costs). The robot experiment in the previous paper (Porr and Wörgötter, 2002) demonstrates that in a real world application already few resonators ($N = 10$) suffice to obtain the desired obstacle avoidance behaviour after learning.

Now we have to consider more complex transfer functions for P_1. Up to this point we have set $P_1 = 1$ which means that the disturbance basically reaches the input X_1 un-filtered which is in general not the case. Due to specific sensor-properties and due to properties in the environment the disturbance reaches the input X_1 in a filtered form. All these changes can be subsumed from the organism's point of view by the function P_1 (and the same applies to P_0). We recall that we have used a Taylor approximation of Eq. 14 and matched it with the sum of resonators to obtain the coefficients. This, however allows concluding that any transfer function P_1 of the shape:

$$P_1 = \frac{(s+z_0)(s+z_0^*)\dots(s+z_n)(s+z_n^*)}{(s+p_0)(s+p_0^*)\dots(s+p_m)(s+p_m^*)} \tag{23}$$

can still (together with the delay term $-e^{-sT}$) be approximated by a sum of resonators, because this sum continues to take the shape of a broken rationale function similar to that in Eq. 23 above[3]. Such a shape of P_1, however, covers all generic combinations of high- and low-pass characteristics. Hence it represents a standard passive transfer function. In addition, we can normally assume that the environment does not actively interfere with signal transmission in such a system and can therefore – with great likelihood – be represented by Eq. 23. Thus, we can argue that an appropriate approximation of the complete Eq. 13 will be found in almost all natural situations. The robot application show in the first paper (Porr and Wörgötter, 2002) supports this notion experimentally.

3.2.3 Convergence Properties

The last section has shown that it is possible to construct approximative solutions of Eq. 13 using resonators so that $X_0(s) \rightarrow 0$. Here we will address the problem if the learning rule will actually converge onto such a solution.

Conventional techniques used to derive a learning rule by calculating the partial derivatives of the weights and finding the minimum fail in our case, because ISO-learning is linear. As a consequence the derivatives are constant and a minimum cannot

[3] Note that we are even able to approximate zero crossings of Eq. 23 since we have a *sum* of resonator responses. If we calculate the overall transfer function of a sum of resonators ($H_1 + H_2 + \ldots$) we automatically get also zero crossings which can be used to identify them with the zero crossings in Eq. 23. Thus, the approximation is correct including also the phase terms.

be found. An approach, which leads to success, however, is to apply perturbation theory instead.

Let us first treat the system very generally without making *a priori* assumptions as to the characteristics of the H_k. In doing so we can employ perturbation analysis with the nice aspect that we will not make any assumption as to the size of the perturbation. Thus, proof of stability against such a perturbation is equivalent to a proof of convergence. For real resonators this will be a little bit different, though, as we will see below.

Let us assume that we have found a set of weights $\rho_k,\ k > 0$ which solves Eq. 13 and we know that the development of the weights follows Eq.5. Now we perturb the system substituting ρ_j in Eq.5 with $\rho_j + \delta\rho_j = \tilde{\rho}_j$. In order to assure stability we must prove that the perturbation is counteracted by the weight change, thus we must solve Eq.5 hoping to find:

$$\Delta\rho_j \sim -\delta\rho_j \tag{24}$$

Note that this would guarantee convergence because we know that μ is small which prevents oscillations.

After some calculations (see Appendix) we arrive at:

$$\Delta\rho_j = \frac{\mu}{2\pi}\int_{-\infty}^{\infty}\sum_{k=1}^{N}\delta\rho_k - i\omega\frac{|X_1|^2 H_k^-}{1-\rho_0 P_0^- H_0^-}H_j^+ d\omega \tag{25}$$

where we use the superscripts $^+$ and $^-$ for the function arguments $+i\omega$ and $-i\omega$. This result is still general in the sense that we are not necessarily dealing with resonator functions. So we are at the moment still free to make some reasonable assumptions about the set of H_k. Let us, thus, assume orthogonality given by:

$$0 = \int_{-\infty}^{\infty} -i\omega \frac{|X_1|^2 H_j^+ H_k^-}{1 - \rho_0 P_0^- H_0^-} d\omega \;\text{ for }\; k \neq j \tag{26}$$

and we get

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} |X_1^+|^2 |H_j^+|^2 \frac{-i\omega}{1 - \rho_0 P_0^- H_0^-} d\omega \tag{27}$$

In order to prove that the integral in the last equation will be negative (assuring convergence) the inner (reflex) loop which is determined by $\rho_0 H_0 P_0$ needs to be considered. Note, that this loop must at least be stable otherwise the system would not be functional to begin with. Now, there is a theoretical result from the literature (Sollecito and Reque, 1981) which supports the notion that the integral in question is negative as long as the stability of $\rho_0 H_0 P_0$ is guaranteed. Let us try to spell this rather general argument out more concretely[4].

By the use of PLANCHEREL'S theorem Stewart (1960) we transfer the integral in Eq. 27 into the time-domain and get:

$$\Delta\rho_j = \mu\delta\rho_j \int_0^{\infty} a_{x*h}(t) f'(t) dt \tag{28}$$

where we call $a_{x*h}(t)$ the autocorrelation function of $x_1(t) * h_j(t)$ which is the inverse transform of $|X_1^+ H_j^+|^2$ ($*$ denotes a convolution). We note that the remaining term in

[4] In the Appendix we will rigorously prove convergence for the important case of unity feedback.

Eq. 27: $\frac{-i\omega}{1-\rho_0 P_0^- H_0^-}$ contains the derivative operator $-i\omega$ in the numerator. Thus, $f'(t)$ in Eq. 28 is the temporal derivative of the impulse response of the inverse transform of $\frac{1}{1-\rho_0 P_0^- H_0^-}$.

Now we must ask what is the most general condition for the reflex loop (defined by $\rho_0 H_0 P_0$) to be stable. For a concrete stability analysis knowledge of P_0 would be required, which can normally not be obtained. We can, however, in general assume that P_0 being an environmental transfer function should again behave passively and follow Eq. 23. Furthermore we know that the environment *delays* the transmission from the motor output to the sensor input. Thus, P_0 must be dominated by a low-pass characteristic without which it would be unstable[5]. As a consequence we can in general state that the fraction $\frac{1}{1-\rho_0 P_0 H_0}$ is dominated by the characteristic of a (non-standard) high-pass. It follows that its derivative has a very high negative value for $t = 0$ (ideally $= -\infty$) and vanishes soon thereafter. The autocorrelation a is positive around $t = 0$. Thus, the integral in question will remain negative as long as the duration of the disturbance D remains short. As an important special case we find that this especially holds if we assume delta-pulse disturbance at $t = 0$, corresponding to $x_1(t) = \delta(t)$.

Thus, for an orthogonal set of H_k, we have found that ISO-learning will converge if P_0 is dominated by a low-pass characteristic and if we use a disturbance D with a short

[5] Note that the unity feedback condition, treated in the Appendix, represents the simplest possible stable reflex loop. Its *low-pass characteristic* is reduced of being a mere delay in this case.

duration.

Finally, we have to prove, that Eq. 27 is zero in the equilibrium state case where the feedback loop is no longer needed. Thus, we have $0 = X_0 = \rho_0 H_0 P_0$ and the denominator becomes one. We get:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} -i\omega |X_1^+|^2 |H_j^+|^2 d\omega \qquad (29)$$

This integral is anti-symmetrical and thus zero as required. In the first article we had in the open-loop condition discussed that the synaptic weights stabilise as soon as we explicitly set $X_0 = 0$ arriving at the same equation (compare Eq. 21 in Porr and Wörgötter 2002). In the closed loop condition used here this is obtained in a natural way as the result of implicitly eliminating the reflex during the learning process.

3.2.4 Matching the theoretical convergence properties to the practical approach

In this section we will now use real resonator functions for H_k and H_j (see Eqs.2–3). Normally the transfer functions of the resonators are not orthogonal, but we will show by numerical integration that the system still behaves properly.

Here we use the unity feedback condition defined in the Appendix in order to be able to work with a concrete example which is initially stable and we get for Eq. 25:

$$\Delta\rho_j = \frac{\mu}{2\pi}\sum_{k=1}^{N}\delta\rho_k \int_{-\infty}^{\infty} \frac{-i\omega H_j^+ H_k^-}{1 - \rho_0 e^{i\omega\tau}} d\omega \qquad (30)$$

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

[Figure 4 about here.]

Fig. 4a shows the numerically obtained results for $\Delta\rho_j$ as defined in Eq. 30 in the case of a perturbation.

We note that the resonators are not orthogonal since we have nearly for all $j \neq k$ non-zero contributions. The system, however, still compensates for perturbations and, thus, converges, for the following reason. First, consider diagram (a), which represents the case of how the system reacts to a perturbation and look at the diagonal. We find that the values of the integral (Eq. 30) are negative on the diagonal. 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, so we have to discuss them and argue why this does not interfere with the compensation process. 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). Most importantly we observe that the value of the integral (Fig. 4a) is substantially smaller than one 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 Eq. 30, which describes all cross-interference terms, we can argue that perturbations will be eliminated. This is true as long as the sum remains below one, which is realistic, given the small and sign-alternating values of the integral surface.

Thus, from this we realize that strict orthogonality as defined in Eq. 26 is not even necessary to assure convergence. This constraint can be relaxed to the constraint that the absolute value of the sum in Eq. 30 (or Eq. 25, respectively) should remain below one. Thus, for all practical purposes we can concentrate on the behaviour of the diagonal elements even without having to employ an orthogonal set of H.

Fig. 4b shows the equilibrium case with $\rho_0 X_0 P_0 = 0$. We note that in this case the integral is zero for $k = j$ which is in accordance with theory: Since we are in the equilibrium we do not expect any weight changes.

4 Discussion

In this article we have focused on finding a mathematically motivated interpretation of the results from feedback loop (self-referential) based ISO-learning. We were able to show that such a system approximates the inverse controller of the reflex. The theoretical results are at some critical points rather nicely linked to the experimental findings shown in the previous article which support the validity of the theory.

Most of the technical aspects, like necessary assumptions (e.g. the "orthogonality problem") have already been discussed in the sections above. Therefore, we will restrict the discussion here to more general problems.

The inverse controller problem belongs to the most famous problems in engineering. Typical solutions are always based on an intrinsic model (a so called "forward model") of the to-be-controlled system. As opposed to this, our approach is model free because it is based on learning. Furthermore, engineered forward models have the central disadvantage that they will fail if something unexpected happens. Thus, control engineers use their forward controllers always only in conjunction with the feedback-loop controller on which the forward model was originally based. The same strategy is pursued in a natural way in our setup. The double-loop structure of Fig. 3 clearly shows that the reflex will again take over if the outer loop fails. As opposed to engineered systems, however, this will lead to a continuation of the learning process such that the system will continue to improve throughout its lifetime.

A frequently addressed problem in biology is the control of voluntary limb movements, for example in the arm-movement models developed by Haruno et al. (2001) and others. These authors also employ forward models (viz. inverse controllers) to address problems of limb control in a mixed model approach (Wolpert and Ghahramani, 2000). The idea that forward models are involved in motor control has been explored for example by Grüsser (1986) who tried to explain the stability of the visual percept during voluntary eye-movements by means of an internal representation of the motor command, which is called “efferent copy”, “corollary discharge” (von Uexküll, 1926). By now clear evidence exists for such a general mechanism, the details of how it is implemented, however, are still under debate. A discussion about this is beyond the scope of this article, but our theoretical results suggest that sequence order learning can provide a method by which forward models can be generically designed (viz. “learned”). It is conceivable that this observation is not restricted to our specific algorithm but also holds for other temporal sequence learning algorithms like TD-learning.

The above models by Wolpert and Ghahramani (2000), Haruno et al. (2001) or others have in common that they use supervised learning schemes, usually TD-learning to learn the forward model. As stated in the introduction the goal of this set of papers is to provide an un-supervised temporal sequence learning algorithm for autonomous behaviour. An organism which is autonomous can not rely on external rewards. Internal rewards are possible but if we treat autonomy seriously then even an internal reward originates in the last instance from a sensor input. Verschure and Voegtlin (1998) used the same paradigm, namely an (Hebb-like) un-supervised learning algorithm together with a reflex as a reference. However, a novel aspect of our work is that we have taken the environment explicitely into account and introduced it as a non-evaluative structure. Thereby the organism only reacts to the relevant parts of the environment's structure and in the theoretical treatment only aspects of the environment have to be taken into account which either establish the reflex loop or which can be used to supersede it. In that sense the organism does not acquire arbitrary sensorial information but instead *useful* information; useful in the sense of helping it to supersede the reflex (von Glasersfeld, 1996).

The definition of autonomy has been based on the aspect of (un-)predictability of behaviour (Ford and Hayes, 1995, p.11). It is interesting to consider how our system fits into this framework. 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 the reflex *re*action. However, the gain of security for the organism will, on the other hand, lead to an increase of uncertainty observed in the environment. What this means can be understood by reconsidering the robot experiment shown in first article (Porr and Wörgötter, 2002): As long as the robot has only its reflex behaviour it is absolutely predictable for an observer. From the moment learning eliminates the reflex the robot's behaviour becomes more and more unpredictable: although the robot solves its goal (obstacle avoidance) it cannot be predicted *how* the robot actually achieves this. It is specifically this duality of certainty versus uncertainty (depending on the point of view of actor versus observer) which is central to the above addressed definitions of autonomy. Such principles are also identified as the basis for the emergence of social behaviour (Luhmann, 1995).

This series of two papers was meant to provide an alternative framework for temporal sequence learning, which by its linear structure provides better access to analytical treatment than the existing techniques. In addition, we believe that the ISO-learning algorithm could have significant commercial potential, because it can in a model-free way solve various inverse controller problems which should be of relevance for different applied control situations.

Two questions immediately arise which should be addressed by future research. 1) Which modifications have to be done to implement an "attraction" case opposed to the shown "avoidance" case? and 2) Is there a way to implement ISO-learning using spiketrains and biophysically modelled neurons? This, however, extends the scope of this article and is the topic for further investigation.

Acknowledgements

We are grateful to Leslie Smith and to the members of the CCCN-seminar for their helpful comments during various stages of this work. This study was supported by grants from SHEFC RDG “INCITE” and by the European funding “ECOVISION”.

5 Appendix

In this appendix we give the detailed equations for the convergence proof and derive the proof in a rigorous way for the so called *unity feedback condition*.

5.1 Detailed Equations

We continue after Eq. 24. We need to define U and V. U is easy:

$$U_j = X_j H_j = \begin{cases} X_0 H_0 & \text{for} \quad j = 0 \\ X_1 H_j & \text{for} \quad j > 0 \end{cases} \tag{31}$$

V is more complicated. From the definition we have:

$$V = \rho_0 X_0 H_0 + X_1 \sum_{k=1}^{N} \rho_k H_k \tag{32}$$

and from above we know (Eq. 6):

$$X_0 = P_0[V + De^{-sT}] \tag{33}$$

Thus we get for V:

$$\begin{aligned} V &= \rho_0 P_0[V + De^{-sT}]H_0 + X_1 \sum_{k=1}^{N} \rho_k H_k & (34) \\ &= \rho_0 P_0 H_0 V + \rho_0 P_0 H_0 D e^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k & (35) \end{aligned}$$

resulting in:

$$V = \frac{\rho_0 P_0 H_0 D e^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k}{1 - \rho_0 P_0 H_0} \tag{36}$$

Substituting $\rho_j \rightarrow \rho_j + \delta\rho_j$ we get:

$$\begin{aligned} \tilde{V} &= \frac{\rho_0 P_0 H_0 D e^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k + X_1 \sum_{k=1}^{N} \delta\rho_k H_k}{1 - \rho_0 P_0 H_0} & (37) \\ &= V + \frac{X_1 \sum_{k=1}^{N} \delta\rho_k H_k}{1 - \rho_0 P_0 H_0} & (38) \end{aligned}$$

Then calculating the weight change is done using Eq. 5:

$$\Delta\tilde{\rho}_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} -i\omega[V^- + \frac{X_1^- \sum_{k=1}^{N} \delta\rho_k H_k^-}{1 - \rho_0 P_0^- H_0^-}] X_1^+ H_j^+ d\omega \tag{39}$$

where we have introduced the abbreviations $^+$ and $^-$ for the function arguments $+i\omega$ and $-i\omega$.

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

$$\Delta\rho_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} \sum_{k=1}^{N} \delta\rho_k - i\omega \frac{|X_1|^2 H_k^-}{1 - \rho_0 P_0^- H_0^-} H_j^+ d\omega \tag{40}$$

where for X_1 we have made use of the fact that for transfer functions in general we can write: $Y^+Y^- = |Y|^2$ and we have reached Eq. 25 of the main text.

5.2 Introducing the *unity feedback loop* restriction

The basic (critical) property of a reflex loop is 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 function, on the other hand, are secondary and can, therefore, be simplified.

Thus, we will use the so-called *unity feedback loop* assumption to capture this property. It is defined by:

$$\rho_0 \in]-1, 0[\qquad (41)$$

$$H_0 := 1 \qquad (42)$$

$$P_0 := e^{-s\tau} \qquad (43)$$

The 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 defined by Eq. 41 results from the demand that the reflex should be a negative feedback loop and that it must be stable.

In addition, we assume that also the transfer function P_1 of the predictive pathway represents un-filtered throughput given by:

$$P_1 := 1 \qquad (44)$$

Finally we assume that the disturbance D should be short with a duration which is shorter than τ (otherwise the loops would become unstable) and that it can be developed into a product series of conjugate zeroes and poles (e.g. low-/band- or high-pass characteristics). Thereby, D also takes on the property of a typical transfer function.

5.3 Convergence for unity feedback

In the main text we had arrived at Eq. 27:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} |X_1^+|^2 |H_j^+|^2 \frac{-i\omega}{1-\rho_0 P_0^- H_0^-} d\omega \tag{45}$$

and we have to prove that this integral is negative.

This can be directly shown for unity feedback. Thus, Eq. 45 turns into:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} \underbrace{|DH_j|^2}_{A(i\omega)} \underbrace{\frac{-i\omega}{1-\rho_0 e^{i\omega\tau}}}_{-i\omega F(-i\omega)} d\omega \tag{46}$$

As in the main text we apply PLANCHEREL'S theorem to Eq. 46 in order to transfer the integral back into the time-domain and prove that it is negative. We get:

$$\Delta\rho_j = \mu\ \delta\rho_j \int_0^{\infty} a(t) f'(t) dt \tag{47}$$

The function $F(s)$ of Eq. 46 is given by the transformation pair:

$$F(s) = \frac{1}{1-\rho_0 e^{-s\tau}} \ \leftrightarrow\ f(t) = (-1)^n \delta(t - n\tau), \quad n = 0, 1, 2, \ldots \tag{48}$$

where f represents an alternating δ-function at $t = 0, \tau, 2\tau, \ldots$ which starts with a positive delta-pulse (Doetsch, 1961). Thus, together with $-i\omega$ the complete term $(-i\omega\frac{1}{1-\rho_0 e^{i\omega\tau}})$ represents $f'(t)$, hence the temporal derivative of f.

The other term $A(s)$ of Eq. 46 is given by:

$$A(s) = |DH_j|^2 \;\leftrightarrow\; a(t) = \Phi[d(t) * h_j(t)] \tag{49}$$

where $*$ denotes a convolution and Φ the autocorrelation-function.

As a consequence of the above findings we have to discuss the integral in Eq. 47 specified by the time functions in Eqs. 48 and 49. The integral should be negative to assure stability. We know that D is short-lived with a duration shorter than τ, without which the loop-system would be instable to begin with. Thus, we can restrict the discussion of the integral to $t = 0$. We know that the autocorrelation function a has a positive maximum at $t = 0$ and that the derivative f' of a delta-pulse at zero approaches $-\infty$ for $t \to 0;\ t > 0$. As a consequence the integral is negative as required for convergence.

References

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

D'Azzo, J. J. (1988). *Linear Control System analysis and design*. Mc Graw, New York.

Doetsch, G. (1961). *Guide to the applications of the Laplace and z-Transforms*. Van Nostad Reinhold Company, London.

Ford, K. M. and Hayes, P. J., editors (1995). *Android Epistemology*. MIT-Press, Cambridge.

Grüsser, O. (1986). Interaction of efferent and afferent signals in visual perception. a history of ideas and experimental paradigms. *Acta Psychol*, 63:3–21.

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

Luhmann, N. (1995). *Social Systems*. Stanford University Press, Stanford, California.

McGillem, C. D. and Cooper, G. R. (1984). *Continous and discrete signal and system analysis*. CBS publishing, New York.

Nise, N. S. (1992). *Control Systems Engineering*. Cummings, New York.

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

Porr, B. and Wörgötter, F. (2002). Isotropic sequence order learning. submitted to Neural Comp.

Sollecito, W. and Reque, S. (1981). Stability. In Fitzgerald, J., editor, *Fundamentals of System Analysis*, chapter 21. Wiley, New York.

Stewart, J. L. (1960). *Fundamentals of signal theory*. Mc Graw-Hill, New York.

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

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

von Glasersfeld, E. (1996). Learning and adaptation in constructivism. In Smith, L., editor, *Critical Readings on Piaget*, pages 22–27. Routledge, London and New York.

von Uexküll, B. J. J. (1926). *Theoretical biology*. Kegan Paul, Trubner, London.

Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience supplement*, 3:1212–1217.

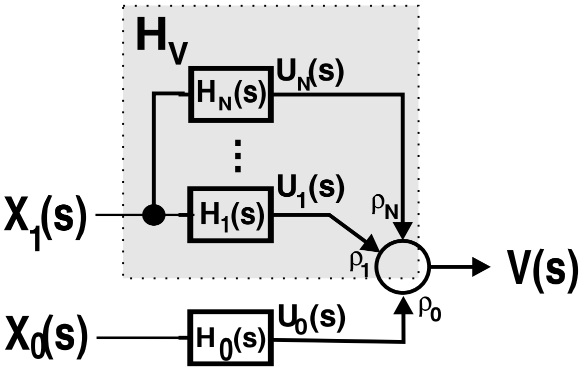

Figure 1: The neuronal circuit in the Laplace domain. The shaded area marks the connections of the weights ρ_k with $k \geq 1$ onto the neuron. The overall contribution from these input is called H_V.

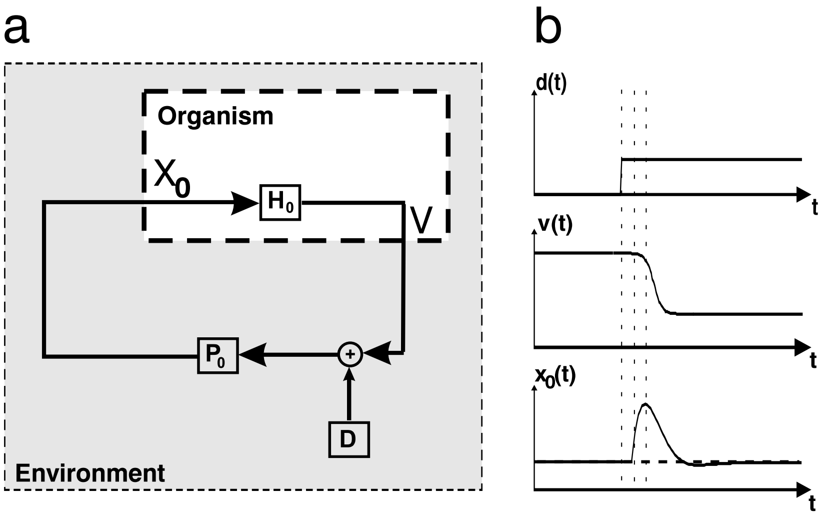

Figure 2: Fixed reflex loop: the organism transfers a sensor event X_0 into a motor response V with the help of the transfer function H_0. The environment turns the motor response V again into a sensor event X_0 with the help of the transfer function P_0. In the environment there exists the disturbance D which adds its signal at $\oplus$ to the reflex loop. ***b)*** *Possible temporal signal shapes occurring in the reflex loop when a disturbance $d \neq 0$ happens. 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 $\oplus$.*

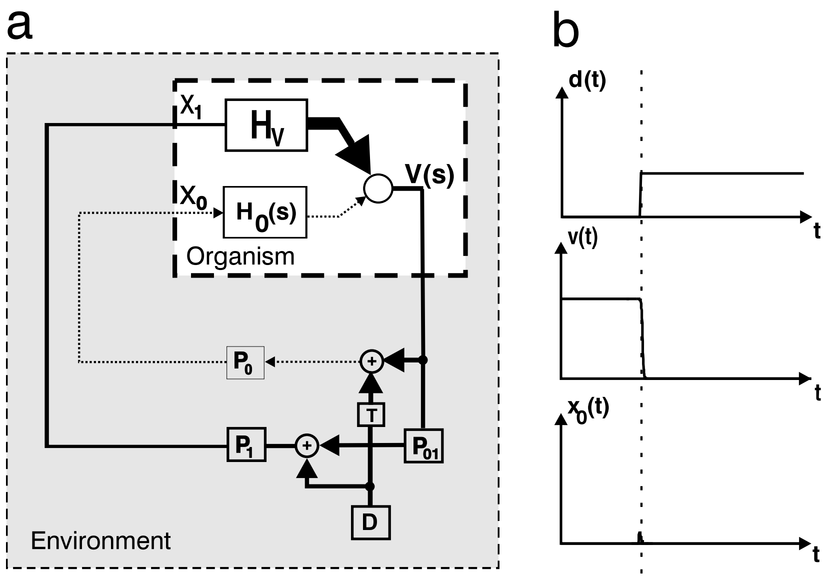

Figure 3: Schematic diagram of the augmented closed loop feedback mechanism which now contains a secondary loop representing ISO-learning. a) H_0 and P_0 form the inner feedback loop already shown in Fig. 2. The new aspect is the input-line X_1 which gets its signal via transfer function P_1 from the disturbance D. The inner feedback loop receives a delayed version (T) of the disturbance D. The adaptive controller H_V has the task to use the signal X_1, which is earlier than and, thus, "predicts" the disturbance D at X_0, to generate an appropriate reaction at V to prevent a change at X_0. b) Shows a schematic timing diagram for the situation after successful learning when a disturbance has occurred. The output v sharply coincides with the disturbance d and prevents a major change at the input x_0.

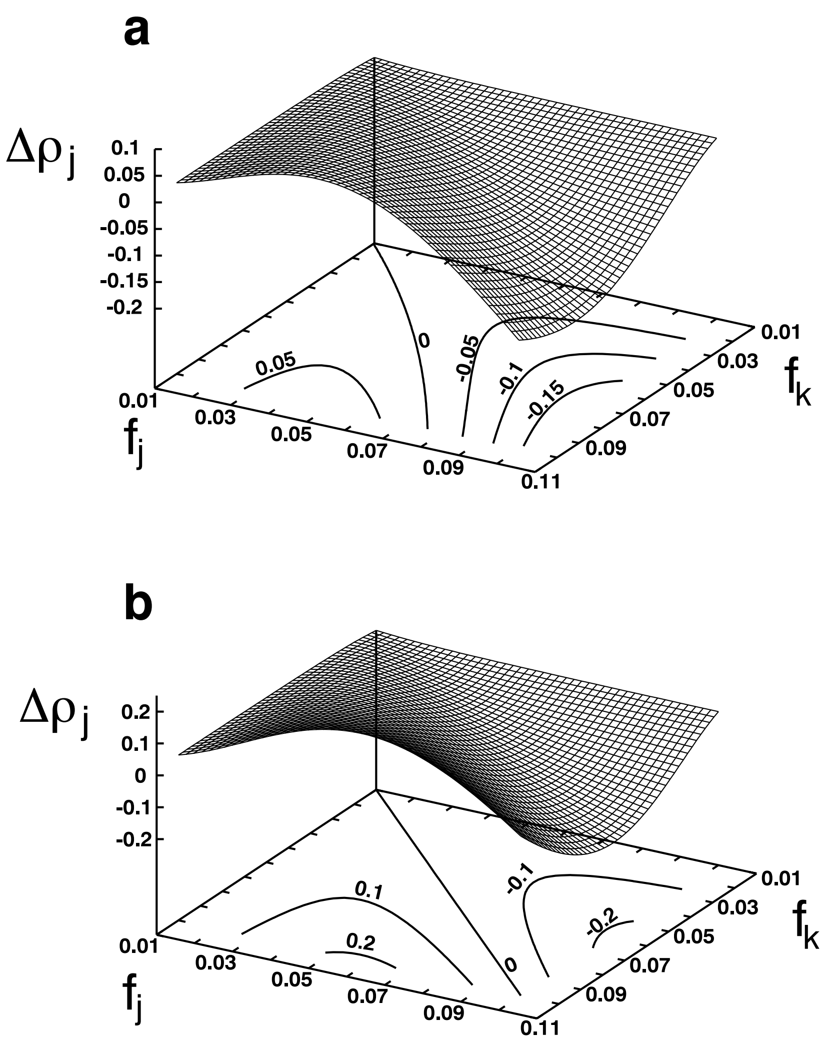

Figure 4: Numerical integration of Eq. 30. The disturbance D and the reflex loop delay τ were both set to one. The frequencies of the resonators H_k and H_j were varied from 0.01 *to* 0.1 *in steps of* 0.001. *The value of Q was set to $Q = 0.9$ for both resonators. The weight of the reflex loop was $\rho_0 = -0.9$.*