UNIVERSITY OF STIRLING

Sequence-Learning in a Self-Referential Closed-Loop Behavioural System

by

Bernd Porr

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Human Sciences
Department of Psychology

May 2003

UNIVERSITY OF STIRLING

ABSTRACT

FACULTY OF HUMAN SCIENCES
DEPARTMENT OF PSYCHOLOGY

Doctor of Philosophy

by Bernd Porr

This thesis focuses on the problem of "autonomous agents". It is assumed that such agents want to be in a desired state which can be assessed by the agent itself when it observes the consequences of its own actions. Therefore the *feedback* from the motor output via the environment to the sensor input is an essential component of such a system. As a consequence an agent is defined in this thesis as a self-referential system which operates within a closed sensor-motor-sensor feedback loop.

The generic situation is that the agent is always prone to unpredictable disturbances which arrive from the outside, i.e. from its environment. These disturbances cause a deviation from the desired state (for example the organism is attacked unexpectedly or the temperature in the environment changes, ...). The simplest mechanism for managing such disturbances in an organism is to employ a reflex loop which essentially establishes reactive behaviour. Reflex loops are directly related to closed loop feedback controllers. Thus, they are robust and they do not need a built-in model of the control situation.

However, reflexes have one main disadvantage, namely that they always occur "too late"; i.e., only *after* a (for example, unpleasant) reflex eliciting sensor event has occurred. This defines an objective problem for the organism. This thesis provides a solution to this problem which is called Isotropic Sequence Order (ISO-) learning. The problem is solved by correlating the primary *reflex* and a predictive sensor *input*: the result is that the system learns the temporal relation between the primary reflex and the earlier sensor input and creates a new predictive reflex. This (new) predictive reflex does not have the disadvantage of the primary reflex, namely of always being too late. As a consequence the agent is able to maintain its desired input-state all the time. In terms of engineering this means that ISO

learning solves the inverse controller problem for the reflex, which is mathematically proven in this thesis. Summarising, this means that the organism starts as a reactive system and learning turns the system into a pro-active system.

It will be demonstrated by a real robot experiment that ISO learning can successfully learn to solve the classical obstacle avoidance task without external intervention (like rewards). In this experiment the robot has to correlate a reflex (retraction *after* collision) with signals of range finders (turn *before* the collision). After successful learning the robot generates a turning reaction before it bumps into an obstacle. Additionally it will be shown that the learning goal of "reflex avoidance" can also, paradoxically, be used to solve an attraction task.

Contents

Acknowledgements

I would like to thank the Psychology department of the University of Stirling for the financial support. Especially I would like to thank Roger Watt and Peter Hancock.

I would also like to thank my supervisors Florentin Wörgötter and Barbara Webb for their support and excellent feedback during all stages of the thesis.

I had very fruitful discussions with Leslie Smith, Colin Grant, David Lieberman, Anders Lansner, Ulf Eysel, Bill Phillips, Matthias Henning, Norbert Krüger, Markus Dahlem, Aušra Sandergine, Richard Reeves, Peter Hancock, Erik Fransén, Shirley A. Plant, Jeannette Hoffmann, Mikael Djurfeldt, Thomas Mittmann, Charlie Frowd, Peter Uhlhaas and in general I always got an excellent feedback from the CCCN-seminar.

Many thanks to Claire Thomson, Euan Foy and Peter Sutherland who did the proof reading.

I would also like to thank Peter Hucker and Steven Stewart in the electronics-workshop for the technical support.

Thanks to Bob Lavery and Bruce Sutherland who supported me in the installation of the video equipment in the robot playground.

A thank you to all the administrative people in the department of psychology, especially Kerry Fairbairns, Kay Bridgeman, Claire Wilson, Cathie Francis and Penny House.

Most importantly: I thank Annette for her endless patience and that she has always supported me during the writing of this thesis.

Declaration

Publications based upon the work contained in this thesis:

Porr, B. and Wörgötter, F. Isotropic Sequence Order Learning in a Closed-Loop Behavioural System. *Proceedings of the Royal Society.* In Press.

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

Porr, B., v.Ferber, C. and Wörgötter, F. (2003). ISO learning approximates a solution to the inverse controller problem in an unsupervised behavioural paradigm. *Neural Computation*, 15:865–884.

Porr, B. and Wörgötter, F. (2003). Learning a forward model of a reflex. *Proceedings for the conference "Neural Information Processing Systems 2002"*, Vancouver, in Press.

Porr, B. and Wörgötter, F. Interaction, self-reference and contingency in computational neuroscience: analytical descriptions and information theoretic consequences. *Proceedings of the conference "Fictions of Dialogue: Interdisciplinary Approaches"*, 23-25 November 2001 in Edinburgh, in press.

Porr, B. and Wörgötter, F. (2002). Isotropic Sequence Order Learning in a Closed Loop Behavioural System. *Proceedings of the EPSRC/BBSRC International Workshop - Biologically-Inspired Robotics: The Legacy of W. Grey Walter 2002* in Bristol, HP technical report.

Porr, B. and Wörgötter, F. (2002) Isotropic Sequence Order Learning using a Novel Linear Algorithm in a Closed Loop Behavioural System. *Biosystems*, 67:1–3, 195–202.

Porr, B. and Wörgötter, F. (2002) Predictive learning in rate-coded neural networks: A theoretical approach towards classical conditioning. *Neurocomputing*, 44–46, 585–590.

Porr, B. and Wörgötter, F. (2001) Temporal Hebbian learning in Rate-Coded Neural Networks: A theoretical approach towards classical conditioning. *Proceedings of the ICANN 2001*, Vienna.

World Patent (pending). Predictive Filter: Controller and Method of Controlling an Apparatus. Filed 05.6.2001, in the name of University of Stirling, B. Porr and F. Wörgötter (Inventors)

Chapter 1

Introduction

1.1 Introductory remarks

The philosophical background of this thesis is constructivism (Maturana and Varela, 1980) and the social theory by Luhmann (1995). The aim of this chapter is to introduce these two theories and make the reader familiar with often non-intuitive consequences.

Constructivism is only the underlying paradigm. The actual focus in this thesis is on *autonomous agents* which will be introduced in section 1.2. In the sections 1.3–1.5 autonomous agents will be discussed in the light of constructivism and Luhmann's system theory.

After having introduced the underlying paradigm and the agent itself section 1.7 will state the central question of this thesis: "How can a reactive agent turn itself into a proactive agent?". Consequently, first reactive behaviour will be introduced, then proactive behaviour. Finally it will be suggested how learning could achieve this.

1.2 Autonomous Agents

Organisms act in their environment. Action shall be understood by any alteration of the environment in a passive or active way. This alteration can be observed by an external observer or by the organism itself. This thesis emphasises the point

of view that the organism is the observer of its own actions. In other words the organism's own perspective is radically employed.

Self observation of its own actions has a certain purpose. Usually the purpose is to determine if actions have changed the environment in the right way (from the organism's perspective). Specifically the organism acts in the environment in order to achieve a desired state. When I touch a hot surface and pull my hand away I have done this in order to reestablish the desired state, namely not to feel pain (any more). In order to achieve this state an appropriate motor reaction has been issued which is suitable to change the relation of the organism to its environment in a desired way. Summarising, the organism has formed together with the environment a closed loop.

The example of a the hot surface made it clear that an organism wants to get into a *desired state*, namely, in the example, that no pain is felt. A state from the organism's perspective can *only* be measured at its inputs and never at its outputs. Therefore it can be stated:

> Organisms control their inputs and not their outputs (von Glasersfeld, 1996)[1].

The above example (hot surface) features a *simple reflex* and illustrates its inherent disadvantage: it always occurs too late. The hot surface first has to be touched and only then the hand can be pulled back. This poses an objective disadvantage of any feedback loop and therefore an objective problem in a very generic way. A solution to this can be found if another sensor event can be found which would *predict* the trigger of the unpleasant stimulus "pain"[2] For example, if we are able use heat radiation as a *predictor* for the trigger of the reflex we can issue an earlier reaction which prevents the trigger of the reflex. Thus, learning the temporal sequence of a predictive sensor event and the sensor input "pain" can eliminate the disadvantage.

---

[1] An external observer would precisely judge this the other way round. An observer would judge that the organism controls its outputs in a way that a specific output state has to be reached. However, only the external observer can see the things like that. If, for example, the action of the organism never feeds back to any sensor input it can not be of any interest to the organism. Therefore a reaction can only be of any interest to the organism when it feeds back to it. This a fundamental difference which will be found throughout this work.

[2] It must be stressed that "pain" means nothing other than the label of an input. In a more strict sense it should be labelled as "reflex input". Therefore labels which involves *interpretations* of sensor signals are written in curly braces in this thesis.

Summarising, there are two aspects which are important for this work. The first aspect is that the *organism is observing itself.* Only the self-observation of the organism shall be of interest here which leads to a self-referential description of the organism which is known as constructivism (Maturana and Varela, 1980). This paradigm shall be used as a basis in this work. The other aspect is the objective disadvantage of feedback loops. It will be shown that learning of the temporal sequence of sensor events can be used to generate new sensor-motor loops which do not have the disadvantage of the original late reacting sensor motor loops. This leads to the field of temporal sequence learning. Thus, this thesis deals with temporal sequence learning in the framework of self-reference and more in general with constructivism. The goal of this work is to develop a self-referential description of temporal sequence learning.

The following paragraphs introduce some aspects of constructivism and sequence learning in more detail which are important to this thesis.

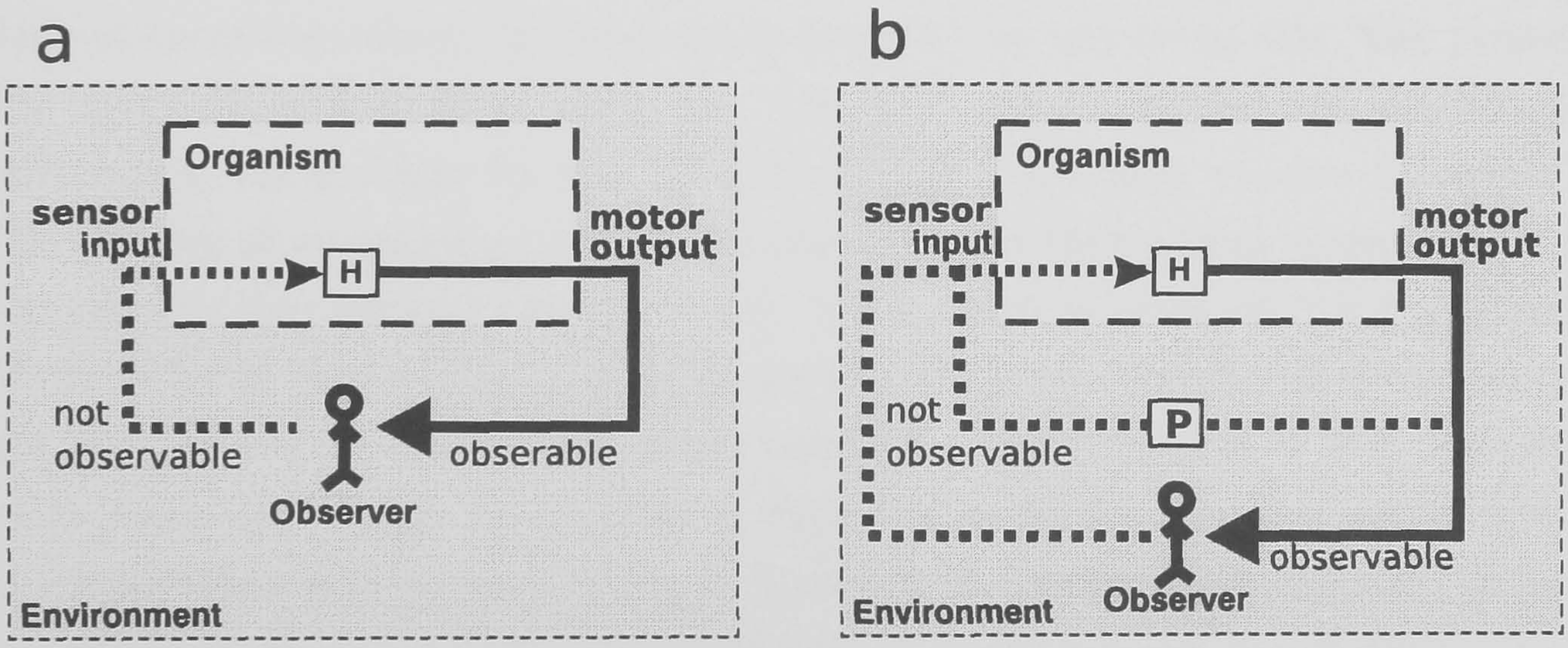

FIGURE 1.1: Observer-problems: the solid lines show observable aspects and the dotted lines show aspects which can not be observed. H transfers a sensor-signal to a motor-reaction. P is the property of the environment and transfers a motor-reaction into a sensor-stimulus.

1.3 Observer-problems

This paragraph introduces *problems* which arise when organisms are *observed* (Luhmann et al., 1990, pp.7–11). The observer has no access to the internal processes of the organism (see Fig. 1.1). Therefore the only observable aspect of the organism is its *behaviour* (see Fig. 1.1a, solid lines).

However, the observer assumes that stimulating the organism's sensors has a partial causal effect on the organism's motor reactions. For example, when I talk to another person I expect that that person will respond to me. The person perceives my sentence and will probably respond with a behaviour — usually with a sentence. However, there is usually no observable direct relation between a stimulus and a response when an organism is observed. There are at least two reasons which makes it difficult for an observer to establish a causal relationship between a stimulus and a response. First, because of internal (or hidden) processes in the organism it becomes difficult to formulate a causal relationship between stimulus and response. An individual ontogenesis of every organism makes it even more difficult to establish causal relations since it becomes more and more difficult to generalise from one organism to another. This is probably the case in nearly all everyday situations where the behaviour of a person is no longer directly explainable by observing another person. Every person has his/her personal history and every person has an extremely complex nervous system. In order to compensate for the lack of knowledge about the observable causality we use terms like "the person has made up his/her mind" or it has "free will" and so on. Second, as already mentioned, the problem for any observer is that it can only observe behaviour. The sensors of an organism can be identified but not their actual operation. One can observe that another person has ears but for such an observer it is not trivial if the auditory information is actually used by the person or not. If the auditory information has an effect on the other person is usually concluded from the person's *behaviour*. A person enjoying a daydream is usually not very aware what is happening in his/her surroundings. Speaking to a person enjoying a daydream usually leads to no *reaction*. From that lack of reaction it can be concluded that the person is not using the auditory stimulus. However, there might be other reasons (the person is deaf, ignorant, ...). Thus, observers are doomed to *interpret behaviour*. This is a very generic observer problem and it cannot be solved as long as the internal processes are hidden inside the organism.

All the above observer-problems have arisen by observing the organism as an input/output- or stimulus/response-system. So far it has not been explicitly mentioned that the organism lives in an environment. For the organism itself the environment has the important aspect of providing feedback from the motor outputs to the sensor inputs. This is usually called a closed sensor motor loop and the task of identifying this loop leads to another observer-problem: Imagine an observer has the task of finding the sensor-motor feedback loops by watching the organism's behaviour (see Fig. 1.1b). As mentioned before the observer can only

observe the *behaviour* of the organism. However, to identify the feedback loops the observer has to identify those sensor-inputs which are able to close the sensor-motor loop. This is a very hard task and it is not very probable that the observer will identify the right feedback in the environment. It is more probable that the observer will identify only those loops which contain the observer him-/herself (see Fig. 1.1b). However, the organism might use (exclusively) feedbacks which do *not* contain the observer (see P in Fig. 1.1b). Imagine a teacher who is teaching a class. The students are surprisingly silent and seem be listening. The teacher interprets this silence as his/her personal success of teaching because he/she can impress the students with his/her charisma. Translated into the current vocabulary used in this thesis the teacher thinks that his/her feedback loops include the students. However, the students are not even listening and are silently playing an exciting card-game under the table which rather integrates the other mates in their feedbacks and not the teacher[3]. Thus, for an observer (here the teacher) it is extremely difficult to identify the right feedback loops since the observer is tempted to integrate him-/herself into the feedback loop of the organism being observed. The organism can use completely different sensor-motor loops which do not contain the observer at all.

However, there is no need to place the observer in the environment of an organism. The organism can be its own observer. In this thesis this shall be the point of view which leads to the next section.

1.4 Self-reference

From now on the organism's perspective shall be emphasised and not the perspective of an external observer. This leads to another interpretation of the feedback loop through the environment. From the organism's point of view only actions which *feed back* to the organism can be of any interest (see Fig. 1.2). Any other action which disappears in the environment can not be of any interest to the organism (von Glasersfeld, 1996) (such actions could only be of interest for an observer)[4]. From this philosophical point of view it is understandable that the organism is only working in *self-contact* with the environment. The role of the

[3]That is probably the reason why pedagogics loves constructivism and Luhmann's system-theory. Students often fool the teacher and there is usually not the linear information transmission from teacher to student and back. This usually remains the dream of the teacher.

[4]In this thesis all evolutionary processes are explicitly excluded. The starting point shall be ontogenesis of an individual.

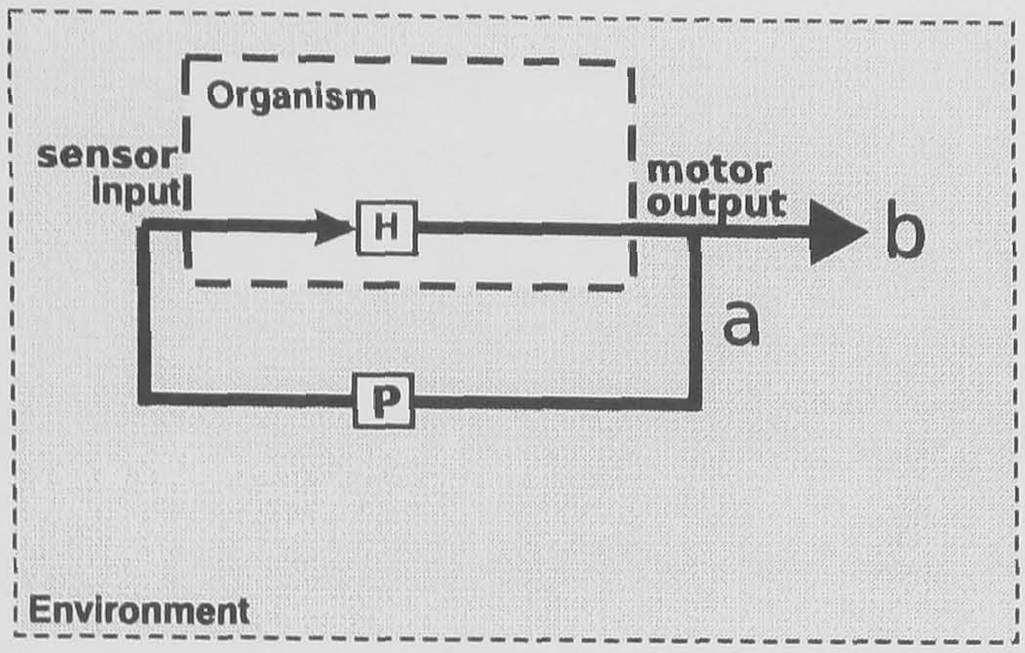

FIGURE 1.2: The organism as an observer: H transfers a sensor-signal to a motor-reaction. P is the property of the environment and transfers a motor-reaction into a sensor-stimulus. The organism as an observer is only interested in those aspects of it own behaviour which feed back to its sensor-inputs (a). Any behaviour which never feeds back can not be of any interest (b).

environment from the organism's point of view is (only) to provide a (maybe poor or noisy) feedback. Therefore it can be stated that the operations of an organism are self-referential operations.

As a consequence the control-paradigm of this thesis will be that of a *closed loop system* opposed to input/output- or stimulus/response paradigms. Self-referentiality is only possible if the organism's perspective is employed and the organism is placed in an environment which provides feedback. Any formulation below must start with closed loops and resulting learning-rules must be treated in this framework.

Closed loops can be stable or unstable. At this point Maturana's principle is employed that states that feedback loops have to "work" (Maturana and Varela, 1980)[5]. This means nothing other than that they have to be stable and have to fulfill a certain *function for* the organism (von Uexküll, 1926). Stability and functionality of a feedback loop are judged by the organism itself and not by its environment. In the above example of the school the students are convinced that their autopoiesis works perfectly fine. However, the teacher probably has another opinion about the card game under the table (and will soon disturb this autopoiesis).

The observer-problem of identifying the right feedback loop can now be discussed from the organism's perspective. The observer has the problem that he/she will

[5]This is a direct result of Maturana's definition of autopoiesis (which implies self-maintenance and internal stability) and it will be used especially at the end of the thesis in the section "embodiment".

often identify the wrong feedback loops. The organism itself cannot develop arbitrary feedback or make errors because in the worst case the organism could even die. In a milder condition the organism would realise that the feedback does not "work" properly and would adjust it accordingly. Therefore, finally for the organism the feedback is always sufficiently adjusted to its situation[6].

Interpreting the organism as a closed loop system leads to a self-referential description. Motor reactions cause sensor-changes and the sensor-changes cause motor-reactions and so on. However, the closed loop itself shall not be sufficient for a strict definition of self-reference like Luhmann (1984, pp.57) has defined it. Luhmann demands that there are no conversions between different qualities. Thus the description must be purely mathematical without the need for labelling any quantity (like 1*meter*, firing rate, ...). However, the environment *does* consist of different qualities like light, pressure and other *physical* quantities. The goal is now to find a description which does not need any transformation from one modality into another. This problem can be solved if the organism's point of view is taken.

Von Uexküll (1926) and von Foerster (1960) argue that at the sensor surfaces of an organism all sensorial qualities are eliminated and converted to neuronal signals. The same applies to the motor output but only the other way round. The *sensor-motor* loop now enables us to describe the organism in a self-referential way by *only* using its neuronal activities. Since the motor output *feeds back* to the sensor surfaces motor signals lead again to sensor signals. As a consequence an organism can be described as a self-referential system which means that neuronal activity leads to neuronal activity and so on (Ashby, 1956; von Foerster, 1985).

The elimination of all modalities leads to a description which only transforms quantities into other quantities. This is nothing more than a pure mathematical description. In such a description a signal is simply transformed into another signal. In addition the mathematical description must be able to deal with the recursivity of a closed loop. Control theory or signal theory which originated in electrical engineering is an appropriate tool since it offers a well developed mathematical toolbox to explore closed loop systems (Stewart, 1960; Sollecito and Reque, 1981; McGillem and Cooper, 1984; D'Azzo, 1988; Terrien, 1992; Nise, 1992; Palm, 2000). These methods shall be used throughout this thesis.

[6]This is one of the basic assumptions in a constructivist therapy. The patient him/herself never has problems generated by him/herself. The problems are always social but never personal. They only become personal because of feedback from the environment (Watzlawick, 1990).

At this point one should keep in mind that the feedback principle has to be seen in very general terms because the organism can change by learning and this might lead to acquisition of additional loops which first have not been taken into account by the organism. Additionally it might happen that the organism no longer uses a certain feedback since it becomes inactive. However, in general only actions which feed back to the sensor surfaces can be of any interest to the organism and this principle will be called *action-feedback*.

1.5 System-levels

In the previous section the perspective of the organism lead to a self-referential description of the organism by neuronal signals. It is obvious that there exist other self-referential systems which employ other *system-levels*. Specifically the self-referential system of behaviour shall be introduced here as another system-level to demonstrate its fundamental difference from the self-referential neuronal system.

It has been pointed out above that for an *external* observer only *behaviour* is observable. Consequently one can concentrate only on behaviour and can take behaviour as a basis for a self-referential system: Behaviour triggers behaviour and so forth. Such a system is called an action-system and anything else is omitted (Parsons, 1951). If one analysed a dialogue in the context of the behavioural (or action-) system one would describe how many reactions are possible following a specific action. Therefore the analysis of social systems leads to the analysis of the *behavioural repertoire* (including language). For example, one could analyse and monitor the discussion of a subject in a group. A measure could be the *complexity* of the discussion measured by the number of the different ways to react.

This example demonstrates that in a behavioural system any attribution towards neuronal or internal states is completely omitted and that the quantitative description is only taken from the observable behaviour. For example, it cannot be said that a person becomes annoyed because of the action of another person. Being annoyed is an internal state and can only be concluded from the observed actions. Therefore in the context of the system-theory one would not say that the person is angry but rather that he/she *looks* annoyed (which is observable). This leaves the question open if the person is "really" annoyed or not. Only the facial expression counts and the reaction to this facial expression.

In the above example it became clear the concentration on one self-referential system avoids the observer-problem. The observer always has the problem of finding out if a person is "really" annoyed or not. In everyday life we usually attribute features or states "into" the other's brain, for example if somebody is "angry" or "nasty". The aim of this thesis is not to question if these metaphors are wrong, right or "really represented" in the brain. This question is simply avoided by using the context of constructivism, namely that these metaphors are still simply explanations of *behaviour* (by behaviour, often language). Thus, the observer-problem is avoided by using a self-referential system which consists only of *behaviour*. Behaviour triggers only behaviour. Any attribution towards internal states is not permitted. The above mentioned observer-problems are solved by concentrating only on behaviour and leaving out the guesswork about internal processes (for example, if the person is lazy, greedy, nasty, evil, good, ...)[7].

In the previous section the observer-problems have been solved by concentrating on the self-referential system of behaviour. However, the observer-problems can also be solved by starting from the other epistemological direction, namely by radically employing the organism's perspective. This leads then to a self-referential description which operates only with signals (behaviours do not make any sense here).

The advantage of the self-referential description is that one has to concentrate on one aspect and all other aspects of life can be left out (Luhmann, 1984). By choosing one level of self-reference (either neuronal signals or behaviour) all other underlying mechanisms are left out. For example, if one chooses the self-referential level of (electrical) neural activity, the chemistry of the cells is ignored. The same applies to the system-level of neural activity: If one concentrates on neural activity anything else can be ignored (for example, behaviour). Thus, one has to decide which level of self-referentiality is taken into account. A mixture of different levels is not allowed.

This separation of the self-referential levels is the basis for Luhmann's system-theory. This thesis will mainly use the self-referential level of neuronal signals. The behavioural level will only play a role in the discussion when robot-robot interactions are discussed.

Thus, Luhmann's system-theory tackles the observer-problem by separating the

[7]This implies a very basic rule in a constructivist therapy. What is analysed is the behaviour of a person towards other persons, for example in a family. The success of the therapy is measured if the behaviour of a patient is regarded as compatible to the social surroundings (Watzlawick, 1990).

self-referential system-levels of neuronal signals and behaviour and makes it possible to concentrate on one aspect of life and to leave out all other aspects.

1.6 Other organisms

That the environment is seen only as a feedback does not neglect other organisms in the environment. In the simplest case there are two organisms: ego and alter (Luhmann, 1984) where it is assumed that ego is observing alter. When ego is observing alter it is only interpreted as a special aspect of ego's *environment* or feedback. This distinction is very important since the goals of alter are also defined internally and therefore not observable by ego. However for ego the external and observable *behaviour* of alter is only relevant and not the achievement of the internal goals of alter. Thus, the justification that the other organisms are just part of the environment arises from the fact that only the *behaviour* of alter is observable and not its internal goals.

1.7 From reactive behaviour to proactive behaviour

After having introduced the general concepts of constructivism implementations of self-referential systems shall now be explored. This section will start with a reactive system which only acts after a disturbance has happened. This apparent disadvantage can be eliminated if the agent turns itself into a proactive system which anticipates the trigger of the unwanted reaction. This anticipation has to be learned by the agent itself and therefore learning rules will be presented which have the potential to solve this task. Thus, this section will elaborate how a reactive agent can turn itself into a proactive system.

1.7.1 Reactive behaviour

The simplest form of self-reference in an autonomous agent is a simple reflex. Simple animals rely on reflexes, for example for walking or for finding food but the reflex is also a behaviour which is found in humans. For example, this behaviour

can be seen when somebody touches a hot surface and then he/she pulls their hand away.

However, the reflex behaviour can be applied to more complex situations if the basic property of a reflex loop is taken into account, namely, that it can only react after a certain sensor event has happened. In other words: a reflex is always too late. Keeping this in mind the reflex behaviour can be generalised to situations where one first has a problem and then pulls him/herself out. For example, a person can change his/her life-style after a heart-attack or can do so before when the he/she becomes aware of an unhealthy life-style. Thus, it is the well known distinction between reactive and proactive behaviour.

Von Uexküll (1926) has already pointed out that an organism is only interested in specific aspects of its environment, namely those aspects which form a closed reflex loop or an action-feedback. Loops define which action can change the organism's sensor inputs in a desired way. Therefore from the organism's point of view the feedback loops have the important property of defining what is the actual "world" for the organism.

This leads to another important aspect of any feedback loop. A feedback loop has a *desired state*. Thus, the important aspect of the feedback loop is that it defines a desired state and therefore the goal is to keep this desired state all the time. In the case of the feedback loop it is simply not possible to keep the desired state all the time since the feedback loop only can react when the organism has left the desired state.

All these aspects regarding reflex loops are related to the field of control theory. In the field of control theory a reflex loop is represented by a fixed feedback loop (Ashby, 1956; 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.

1.7.2 Contingency

In control theory noise only plays an implicit role in the sense that it disturbs the control loops. It is the power of classical feedback-control (like PID-controllers) that it works without knowing the explicit origins of the disturbances (Phillips,

2000). However, this thesis will take the noise from the environment explicitly into account.

Noise originating from the environment is seen from the organism's point of view as unexpected events which here shall be called "contingency". Practically this is introduced by a disturbance in the environment which shall be called "D" from now on. This disturbance is again described from the organism's point of view: although there are an infinite number of disturbances in the world, only those disturbances can be of any interest to the organism which actually disturb the *feedback loop(s)*. Since the feedback loop(s) is(are) described in terms of neuronal signals the disturbance can also be described by the organism's internal neuronal signals.

As pointed out above, the feedback loop has the inherent disadvantage that it is always too late. Including a disturbance this can be formulated more precisely: Any feedback loop (or reflex) has the inherent disadvantage that the organism can not predict when a disturbance D will actually happen. As a consequence any organism which relies only on feedback-mechanisms has to cope with unpredictable events from the environment and has to live with the disadvantage that its desired state(s) can not be maintained all the time.

1.7.3 Anticipation

The inherent delay of any reflex behaviour poses an objective problem which has to be solved. This can be achieved if the organism can turn the contingency of D into certainty. This is the case if the organism is able to *predict* the disturbance D and generate an appropriate motor reaction before the disturbance reaches the organism. Again the reflex is the starting-point: The reflex itself can not prevent the sensor event "pain" occurring since it can react only after it has occurred. Only if the organism is able to learn the relation between the "pain" and, for example, the sensation of heat radiation (which precedes it) can it avoid the painful stimulus by generating an anticipatory motor reaction. As heat radiation and pain follow in a *sequence*, learning has the task of learning this temporal sequence to generate an earlier motor reaction.

1.7.4 Temporal sequence learning in a closed loop

The previous paragraph elaborated on the fact that a feedback loop is always too late and that it is therefore not able to maintain a desired state continuously. Thus, the task is to find a learning algorithm which is able to *predict* the unwanted reflex-reaction and which issues a reaction so that the organism's *input* will then *always* be in its desired state.

Learning-algorithms taken from the class of temporal sequence learning are obviously candidates which could eliminate the disadvantage of the feedback- or reflex loop. Temporal sequence learning enables the organism to build up anticipatory structures, to predict looming disturbances and to generate suitable motor reactions to prevent them. Thus, it is necessary to concentrate on the different learning paradigms of sequence learning which are offered in computational neuroscience and biology to decide if one can be used in the closed loop paradigm presented here. A learning algorithm is needed that is able to learn sequences of events and is able to generate appropriate motor reactions.

Learning of sequences has a long tradition in psychology which began with Pawlow's classical conditioning-experiments (Pavlov, 1927). In the classical experiment by Pavlov a dog learns the temporal relation between the food (late event) and the bell (earlier event). A learning-rule which has been inspired by Pavlov's experiments in the field of computational neuroscience is the so called temporal difference learning-rule (TD-learning) which plays a dominant role in many theoretical studies (Sutton, 1988; Montague et al., 1993; Dayan et al., 2000; Haruno et al., 2001; Schultz and Suri, 2001). Here the "later event" is represented by a designated reward- (or punishment-) signal to which the prediction of the learner is explicitly compared. Thus, the reward-signal represents an explicitly defined evaluative feedback for the learning. Learning (weight-change) stops when prediction (the output of TD-learning) and reward match. Obviously the learning scheme by Sutton and Barto is a evaluative learning scheme which needs an external teacher in form of a reward signal.

In the context of constructivism observer-problems arise when an external reward is introduced. First, a reward can only be defined by an observer. However, as pointed out previously, the observer is not aware of the goals hidden in the organism. Therefore it is not probable that the observer gives rewards which are beneficial to the organism. It is much more probable that the observer gives rewards which are beneficial for the observer him/herself. Even if the observer

is of the opinion that he/she gives the rewards for the benefit of the organism it is not clear for the observer if the rewards have been "really" beneficial for the organism and therefore have been rewards at all.

To test if an organism has benefited from a reward the observer can only try to interpret which *behaviours* most resemble the experience of a reward. Alternatively the observer can use his/her own introspection to conclude what has been a reward. All these observations stay on the level of behaviour. TD-learning, however, needs a *signal* which represents a reward. The mapping of the internal (reward-) signal to a behaviour (which looks like a reward) is not permitted in constructivism as it leads to observer-problems (see above). Therefore this thesis can not use a learning rule which relies on teacher-like evaluation.

A learning rule is needed which is non-evaluative in the sense that it does not need a reward signal. This leads to another class of learning rules which are called *unsupervised* learning rules. Amongst these is one learning rule which is of special interest in this context since it learns temporal sequences and is biologically related. New results from neurophysiological experiments suggest that the temporal timing of neuronal signals is crucial to synaptic learning and therefore to synaptic weight change: if the presynaptic activity precedes the postsynaptic activity then the synaptic weight is increased and if the timing is reversed it is decreased. This rule is called spike timing dependent synaptic plasticity (STDP) or simply "temporal Hebb" since it is a special form of classical associative Hebbian learning (Markram et al., 1997; Zhang et al., 1998; Bi and Poo, 2001). While standard Hebbian learning (Hebb, 1967) only develops associations between events which occur at the same time temporal Hebb learns associations between sequences of events. The learning rule operates unsupervised and, thus, seems to be good for explaining autonomous behaviour of an organism since it leads to self-organising (or autonomous) behaviour.

This thesis will use the main features of spike timing dependent plasticity, namely that the weight changes depend on the temporal order of pre- and post-synaptic activity and that learning is correlation-based. The neuronal activity itself is represented by analogue signals which can be interpreted as the firing rate of a neuron.

At first glance it seems to be the wrong way to develop a learning rule in the context of *rate-codes* if the timing of *spikes* is crucial for learning behaviour. However, rate codes also make it is possible to develop learning rules which analyse the timing of pre- and postsynaptic activity. Rate codes have the advantage that

the mathematical framework of signal- and control theory can be used. The link between rate-codes and spike-timing dependent plasticity has been established by Roberts (1999) and also by Xie and Seung (2000). They have proven that one can use a rate-code if the learning rule contains the *derivative* of postsynaptic firing-rates. All rules which operate with rate codes and employ a derivative of the postsynaptic activity are called *differential Hebbian* learning rules since they use the change of the firing rate at the output of the neuron. Differential Hebbian learning can also be divided into supervised and un-supervised learning rules. The above mentioned TD-learning employs also the derivative in its learning rule and therefore belongs to the class of supervised differential Hebbian learning. However, there is a group of differential Hebbian learning rules which operate un-supervised (Sutton and Barto, 1981; Klopf, 1986; Kosco, 1986). These rules are candidates which can be used in the context of constructivism (and this thesis) since they do not use any reward-signal. Additionally these rules operate with analogue signals and can be treated by signal- or control-theory.

However, none of the above mentioned learning rules are designed for the closed loop case. Rather they are designed for the open-loop case and can only be evaluated by an external observer. The important difference between the open-loop case and the closed loop case is the learning goal. In the open-loop condition the *output* has to meet a certain condition. In the closed loop case the *input* has to meet a certain condition ("desired state").

The closed loop condition can be illustrated by the example which describes the task of avoiding a hot surface. The desired state or condition is defined at the sensor-*inputs* of the organism: the "pain"-sensor should always be silent. The organism is interested in the *result* of the action rather than in the action itself[8]. It is clear that a motor-reaction is issued but the motor-action is issued for the *purpose* that the "pain" is no longer felt. In the case of the hot surface it could be a reversal of the motion towards the hot surface or it could be something more sophisticated, for example throwing a cover over it. Therefore usually there is more than one possible reaction which ends the stimulation of the sensor-input "pain"[9].

[8]Even if the organism is interested in the action itself (unity feedback) it can evaluate it only at its inputs.

The same applies to a technical system, for example, central heating. The heater is switched on to get a desired room-temperature. The output of the heater (heat-flow) itself is irrelevant. The only relevant factor is that the heater is able to control the room temperature.

[9]Recent results from classical eye-blink experiments show that conditioned responses and the unconditioned responses are not similar (personal communication with Mikael Djurfeldt of KTH in Stockholm). The rabbits employ a certain form of "laziness" in their conditioned response. From the moment the CS (sound) is felt they close their eyes slowly until the moment the US (air-puff) arrives. Thus, the response is quite similar to the responses which are generated by

This example shows that the *result* of an action is evaluated by the organism and not the action itself.

Learning simply continues from that point of view and also determines its success at the input of the organism. If learning is able to silence the "pain"-sensor all the time it has been successful and learning has fulfilled its task. In the case of the hot surface learning leads to the effect that the motor reaction (namely pulling the hand away) is issued already at the moment when the predicting stimulus (heat-radiation) is felt. Therefore a learning rule is needed which issues an (motor-) action and evaluates the result at the organism's (input-) sensors.

In contrast to the above closed loop case, the goal of the open loop case is usually defined by the learned *reaction* or at the output. From the moment the learned reaction has a similar strength the goal has been reached (Rescorla and Wagner, 1972). For example, in Pavlov's experiment the goal has been reached the moment the amount of saliva caused by the bell and the food is the same. If the amount caused by the bell has reached the same amount caused by the food then learning has reached its goal.

Summarising, in the closed loop condition the learning goal is not defined at the (motor-) output, it is rather defined as a specific *input-condition* (desired state). The observer in this case is the organism itself and the organism observes if a motor-reaction has caused a certain desired effect which is measured at the sensor-surfaces of the organism or in other words: at its *inputs*. This makes clear that it is not the reaction itself (like its strength) that is important to the organism but the result and the result is measured at the *input* as a deviation from the desired state. Therefore closed loop systems control their inputs and not their outputs (von Glasersfeld, 1996).

All the unsupervised learning rules which have already been mentioned use the open-loop paradigm and therefore control their *output* and *not* their *input*. Thus, an *un-supervised (or drive-reinforcement-) learning rule is needed which controls its input and not its output*. Such a learning rule will be presented in this work and will be called *isotropic sequence order learning* or ISO learning.

ISO learning which will be discussed later on. Without going deeper into the subject of air-puff experiments there seems to be evidence that they can be interpreted in both the open-loop-paradigm and in the closed loop paradigm.

1.7.5 The reflex as the boundary condition

The outstanding feature of an un-supervised learning rule, namely that it is self-organising is also its curse: Self-organisation always has the inherent danger that the results become arbitrary and therefore useless to the organism. The standard solution of the theory of neural networks is that so called "boundary conditions" are introduced which reduce the degrees of freedom, so that the network becomes constricted within sensible boundaries. A good example of the application of boundary conditions in classical Hebbian learning is the development of orientation columns in the primary visual cortex of the cat (Miller, 1996a). With the help of boundary conditions it is possible, for example, to tune the size of the orientation columns. The same applies to Linsker's info-max network (Linsker, 1988). There the boundary conditions tune the shapes of the receptive fields. However, these boundary conditions actually only camouflage the experimenter outside the organism who actively interferes preventing the network from becoming arbitrary. Thus, it seems to be that purely unsupervised learning is not applicable and it is clear that some form of reference must exist.

In the autonomous organism of this thesis the solution of preventing its behaviour from becoming arbitrary is the reflex. The reflex is fixed and pre-wired and it can be seen as the "genetic" basis which guides learning. The reflex automatically defines an internal learning goal for the organism which originates from the above stated fact that the reflex always occurs too late. Or more generally: reactive behaviour is always too late and therefore it has to be predicted. Every sensor signal which arrives earlier than the reflex-inducing signal is beneficial to the organism in the sense of being able to predict the unwanted reflex. Any sensor signal on the other hand which comes later is useless.

It is important to mention that the above definition of the learning goal includes only neuronal signals and is therefore absolutely free of any external attribution. The learning behaviour directly originates from the inherent properties of the feedback loops. These properties originate from the causality of time. Therefore the whole learning process can be described in relating neuronal activity with neuronal activity and there is no need to attribute learning goals from outside into the organism like rewards or other evaluations.

1.8 Structure of the following chapters

This thesis discusses an organism in its environment. Thus, there is an organism with sensors and motor-outputs and there is the surrounding environment which closes the loop by providing feedback from the motor-output to the sensors of the organism. The overall structure of the thesis is guided by the observation that the organism is embedded in an environment. Therefore, first only the organism is described and then the organism within its environment is described. However, this division has been done for the purpose of structuring but it does not imply that an organism without environment makes any sense. Even the chapters which focus only on the organism develop an organism which always operates in a closed loop established by the environment.

More specifically chapter 2 will develop the internal structure of the organism while omitting its environment. How the sensor signals are transformed into motor reactions will be presented in section 2.2. As pointed out in the introduction the internal structure of the organism can be changed by a temporal sequence learning what is called ISO learning (section 2.3. Its linear structure allows an analytical treatment of some of its main characteristics (section 2.4). More complex aspects will be addressed by simulations (section 2.5). Thus, chapter 2 will present all aspects of the organism and ISO learning which do not necessarily need the closed loop.

After chapter 2 has treated all aspects of ISO learning which does not need an environment chapter 3 will derive results which need an environment. By embedding the organism in an environment a closed loop situation will be established. This closed loop situation will initially only be established by a simple reflex. The properties of the reflex will be presented in section 3.2. Particularly the lateness of the reflex-reaction will define the goal of the following section, namely to replace the reflex with a faster anticipatory (re)action (section 3.3). It will be shown analytically by applying methods from control theory and perturbation analysis that such a closed loop system creates — by means of ISO learning — a "forward model" of the reflex.

Chapter 4 will support the theoretical findings by a real robot experiment (avoidance case, section 4.2) and by a computer simulation (attraction case, section 4.3). The aim of this chapter is to show the robustness of ISO learning. In addition, to demonstrate that it is possible to establish both an avoidance behaviour and an attraction behaviour out of the same learning rule only be changing the initial

reflex.

The discussion is again guided by the organism and its environment. In chapter 5 only ISO learning is discussed without the surrounding environment. This discussion starts with technical aspects (section 5.2) and then discusses links to neurophysiology (section 5.3). However, the emphasis lies on animal learning (section 5.4) and its mathematical models (section 5.4.5).

Chapter 6 will discuss ISO learning in the context of closed loop learning. The first part of the chapter is mainly devoted to applications in the field of engineering (section 6.2). The second part will discuss indirect consequences of the closed loop paradigm (section 6.3). In particular, observer problems will be discussed. Finally robotics will be discussed as the "natural" closed loop application.

Chapter 2

The Organism

2.1 Introduction

In the last chapter an organism has been introduced which first acts reactively and then, after learning, is able to act pro-actively. To achive this a learning rule has been proposed which performs sequence learning and measures its success at its inputs. The aim of this chapter is to formalise the demands of the last chapter so that in conclusion a mathematical description is at hand.

The arguments of the preceding chapter can be summarised as follows:

- The organism transfers sensor inputs into motor reactions.
- Initially there must be a strong (or maybe fixed) connection between a specific sensor input and the motor output in order to establish a reflex reaction.
- The learning rule must be non-evaluative and allow for learning the temporal correlation between the reflex reaction and other predicting sensor inputs. This temporal correlation should be used to generate an earlier motor reaction to override the reflex.
- The learning goal shall be determined at the inputs of the learning circuit. Learning shall stop if all reflex-inputs have been eliminated.

Using these properties it is now possible to formulate a mathematical framework for the organism.

The following sections will then proceed as follows: first the organism will be mathematically formalised (section 2.2). The resulting equations will define the organism's reactions to sensor-events. Second, a learning rule will be introduced which is able to learn sequences of events and which evaluates its success at its inputs (section 2.3). Third, analytical results will be obtained which show the organism's ability to learn sequences of events (section 2.4) and which will prove that learning determines its success at the input (section 2.4.2). Fourth, simulations will support the analytical findings and will also provide results for cases which are not analytically treatable (section 2.5).

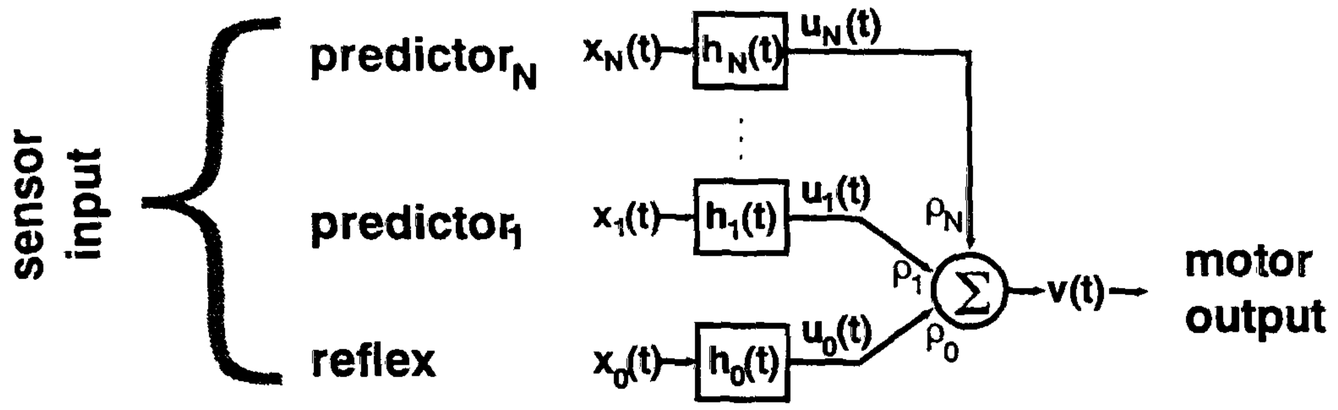

FIGURE 2.1: The basic circuit in the time domain.

2.2 The organism

A system of $N+1$ linear filters $h(t)$ is considered receiving inputs x and producing outputs u. The filters connect with corresponding weights ρ_k to one output unit $v(t)$ (Fig. 2.1).

All input lines of the algorithm presented here are mathematically equivalent. However, h_0 (and the corresponding input x_0) will be used to denote the one unit which will later represent the reflex pathway. The output v is then given as:

$$v(t) = \rho_0 u_0(t) + \sum_{k=1}^{N} \rho_k u_k(t) \tag{2.1}$$

In general, the system which is considered shall operate in continuous time (e.g. with neuronal rate codes) and it shall be able to handle continuous input functions $x(t)$ of arbitrary shape.

The transfer function h shall be that of a *bandpass* which transforms a δ-pulse

input into a damped oscillation (Fig. 2.2a) and is specified by:

$$h(t) = \frac{1}{b}e^{at}\sin(bt) \qquad \leftrightarrow \tag{2.2}$$

$$H(s) = \frac{1}{(s+p)(s+p^*)} \tag{2.3}$$

where p^* represents the complex conjugate of the pole $p = a+ib$. It is important to note that such a bandpass filter is only stable if its pole-pair is located on the left complex half-plane, otherwise an amplified oscillation is obtained. $H(s)$ represents the Laplace-notation. In general low-case letters are used for the time-domain and upper-case letters are used for the corresponding Laplace-transform.

Real and imaginary parts of the poles are given by

$$a = \mathrm{Re}(p) = -\pi f/Q \tag{2.4}$$

$$b = \mathrm{Im}(p) = \sqrt{(2\pi f)^2 - a^2} \tag{2.5}$$

where f is the frequency of the oscillation. The damping characteristic of the resonator is reflected by $Q > 0.5$. Small values of Q lead to a strong damping.

The use of resonators (band-pass filters) is motivated by biology because oscillatory neuronal responses (Traub, 1999) and band-pass filtered response characteristics (at virtually all sensory front-ends, cell-membranes and ion-channels like NMDA) are very prevalent in neuronal systems (Shepherd, 1990). Several examples for the utilisation of such bandpass filtered responses provide Grossberg and Schmajuk (1989) with their spectral timing model which has been used in different applications (Grossberg, 1995; Grossberg and Merrill, 1996).

Thus, the main idea is to use a neuron which gets bandpass filtered sensor signals at its inputs and generates a motor output. Later, one of these band-passes (h_0) has the special task to provide the input for a reflex like reaction. The other bandpass filtered sensor signals are candidates for generating an earlier motor reaction through learning.

2.3 The learning rule

Learning (weight change) takes place according to a Hebb-like rule:

$$\frac{d}{dt}\rho_j(t) = \mu u_j(t)v'(t) \qquad \mu \ll 1 \tag{2.6}$$

where the weight change depends on the correlation between u_j and the derivative of v. All weights can change (also ρ_0). The constant μ is adjusted such that all weight changes occur on a much longer time scale (i.e., very slowly) as compared to the decay of the responses u. Thereby the system operates in the steady state condition.

2.4 Analytical findings

2.4.1 Timing dependence of weight change

In this section the question will be addressed how the timing between the input signals influences the weight change.

To perform analytical calculations two restrictions will be introduced, which will now be used often throughout the theoretical parts of this thesis. They will be waived later:

i) Only two resonators are considered, thus, $N = 1$.

ii) Accordingly the analytical derivation has to deal with only two input functions x_0, x_1 defined as (delayed) δ-pulses:

$$x_0(t) = \delta(t - T), \qquad T \geq 0 \tag{2.7}$$

$$x_1(t) = \delta(t) \tag{2.8}$$

The first restriction is necessary because the analytical treatment of the case $N > 1$ is very intricate and largely impossible.

Concerning the second restriction it must be noted that the theory of signal decomposition allows composing any causal input function from δ-pulses. Thus, the second constraint is not really a restriction.

The delay T assures a well-defined causal relation between both inputs, where x_0 (the later of the two) is the timing reference (the reflex input). Especially the section on the robot implementation will show that the algorithm (with $N > 1$) is very robust with respect to variations in T.

In general as an initial condition will be used:

$$\rho_0 = 1 \tag{2.9}$$

$$\rho_1 = 0 \tag{2.10}$$

For the analytical treatment only the weight change at ρ_1 will be considered. (In fact, a little later it will be shown that the algorithm normally operates always in a domain where ρ_0 changes very little.)

Because steady-state is assumed, the product in the learning rule (Eq. 2.6) can be rewritten as a correlation integral between input and output:

$$\rho_1 \rightarrow \rho_1 + \Delta\rho_1 \tag{2.11}$$

$$\Delta\rho_1(T) = \mu \int_0^\infty u_1(T+\tau)v'(\tau)d\tau \tag{2.12}$$

Similar to other approaches (Oja, 1982) the weight change is computed for the initial development of the weights as soon as learning starts, because this is indicative of the continuation of the learning. Since the weight-change happens on a much slower time-scale than the resonator-responses it can be assumed that Eq. 2.10 not only holds for $t = 0$ but also for $t > 0$ during the first correlation between u_1 and v'. The duration of the correlation is determined by the wavelengths of the resonators and their damping factors and is roughly $t_{\text{response}} = Q/f$. The resonator with the longest temporal response to a delta pulse should be taken as the duration the correlation takes place. This time shall be called t_{corr}. Therefore the assumption:

$$\rho_1(t) = 0 \qquad \text{for} \qquad t < t_{\text{corr}} \tag{2.13}$$

is introduced which reflects the condition during the first pairing of two delta pulses. Assuming that the weight ρ_1 stays zero means that the postsynaptic contribution only originates from the input x_0. Therefore it is possible to replace v' in Eq. 2.12 directly by u_0' and Eq. 2.12 turns into:

$$\rho_1(T)_{t<t_{\text{corr}}} = \mu \int_0^\infty u_1(T+\tau)u_0'(\tau)d\tau \tag{2.14}$$

Thus, Eq. 2.14 calculates the change of the weight ρ_1 under the condition that ρ_1 stays zero.

In simple cases (e.g., for $h_0 = h_1$) this integral can be solved directly. A general solution, which can also be extended to cover more than two inputs, requires to

apply the Laplace-transform using the notational convention: $x(t) \leftrightarrow X(s)$, for a transformation pair of functions in the time and the Laplace domain.

The linearity of the integral Eq. 2.14 allows for an analytical solution, which is possible with the help of Plancherel's theorem (see the Appendix A for this rather little known theorem). Applying it together with the shift theorem $x(t - t_0) \rightarrow X(s)e^{-t_0 s}$ to Eq. 2.14 gives:

$$\Delta\rho_1 = \mu\frac{1}{2\pi}\int_{-\infty}^{+\infty} H_1(-i\omega)\left[i\omega e^{-Ti\omega}H_0(i\omega)\right]d\omega \quad (2.15)$$

$$= \mu\frac{1}{2\pi}\int_{-\infty}^{+\infty} H_1(i\omega)\left[-i\omega e^{Ti\omega}H_0(-i\omega)\right]d\omega \quad (2.16)$$

Note that symmetry of Plancherel's theorem is broken because of the exponential term. Equation 2.15 represents a Fourier transform and Eq. 2.16 an inverse Fourier transform. Note, that these two Equations can be interpreted in these two different ways. This does not mean that in this case the Fourier transform is equal to its inverse. In fact Eq. 2.15 and Eq. 2.16 calculate Fourier transforms and inverse transforms from different functions as the signs swap in the transfer functions H_0 and H_1 when these two equations are compared.

Both integrals can be evaluated with the method of residuals. Eq. 2.16, however, offers the advantage that the right complex half plane can be neglected, because it leads to contributions for negative time (i.e. $t < 0$) only (McGillem and Cooper, 1984; Stewart, 1960). Thus, of the four residuals (poles) for H_1 and H_0 only those of H_1 need to be considered because those of H_0 have flipped their sign in Eq. 2.16 and appear now on the right complex half-plane. We get as the final result:

$$\rho_1(T)_{t=0} = \mu\frac{b_1 M\cos(b_1 T)+(a_1 P+2a_0|p_1|^2)\sin(b_1 T)}{b_1(P+2a_1a_0+2b_1b_0)(P+2a_1a_0-2b_1b_0)}e^{-Ta_1} \quad T \geq 0 \quad (2.17)$$

$$\rho_1(T)_{t=0} = \mu\frac{b_0 M\cos(b_0 T)+(a_0 P+2a_1|p_0|^2)\sin(b_0 T)}{b_0(P+2a_0a_1+2b_0b_1)(P+2a_0a_1-2b_0b_1)}e^{-Ta_1} \quad T < 0 \quad (2.18)$$

where $M = |p_1|^2 - |p_0|^2$ and $P = |p_1|^2 + |p_0|^2$. If identical resonators $H_0 = H_1 = H$ are assumed, this leads to:

$$\Delta\rho_1(T)_{t=0} = \mu\frac{1}{4ab}\sin(bT)e^{-aT} \quad (2.19)$$

which is identical to the impulse response of the resonator itself apart from a different scaling factor.

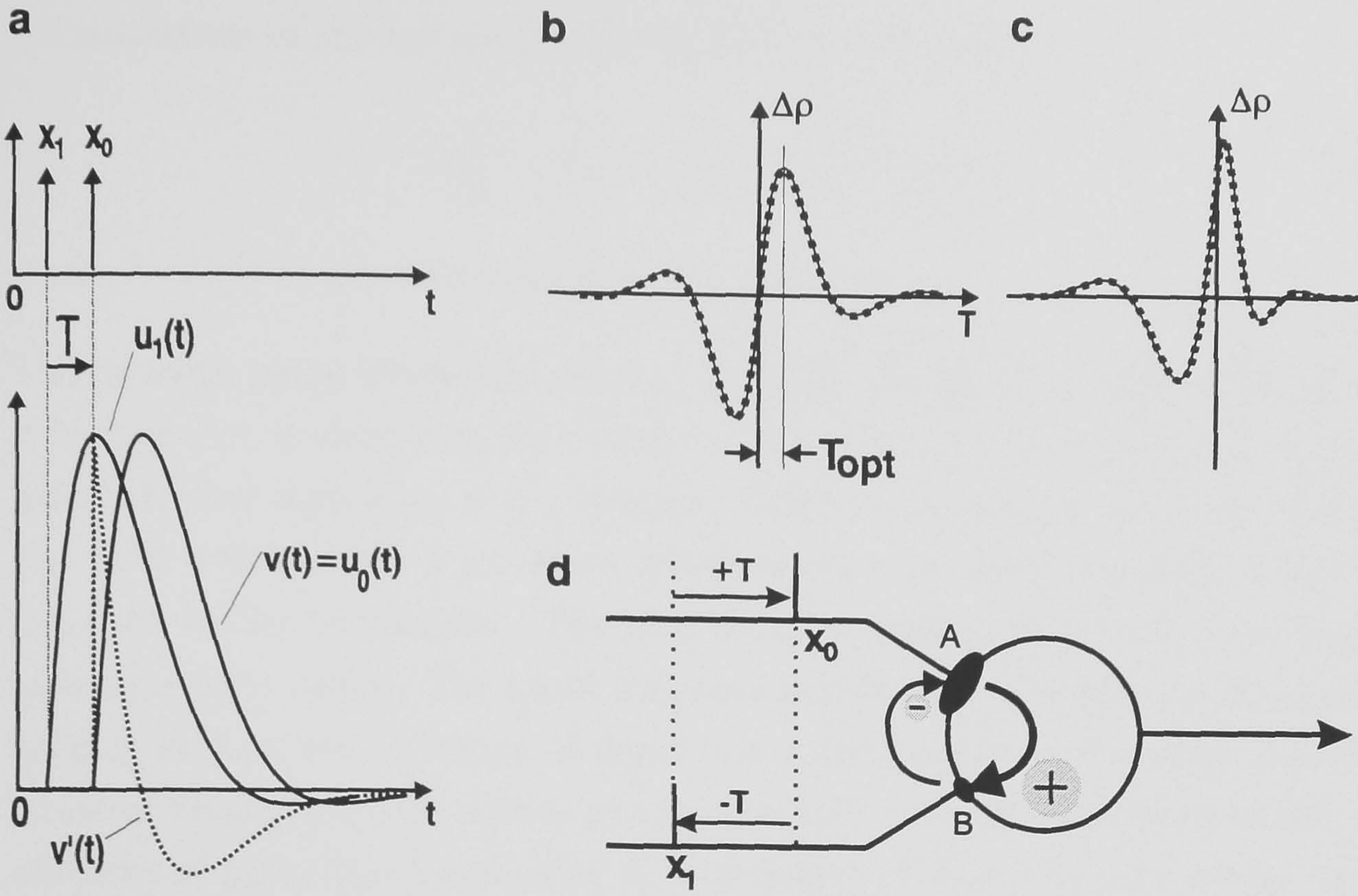

FIGURE 2.2: Input functions and the initial weight change for $t = 0$ according to Eqs. 2.17 and Eqs. 2.18. (a) shows the inputs x, the impulse responses u for a choice of two different resonators h and the derivative of the output v'. (b) shows the initial weight change $\rho_1(T)_{t=0}$ for $H_1 = H_0,\ Q = 1,\ f = 0.01$ (arbitrary units) and (c) for resonators with different frequencies $f_0 = 0.01,\ f_1 = 0.02$ but with the same $Q = 1$. The solid lines in (b) and (c) represent the analytical solutions derived from Eqs. 2.17/2.18 and the dotted lines simulation results resulting from the numerical integration of Eq. 2.12 with the same parameters for f and Q. For that purpose the two filters H_0 and H_1 get two different inputs $x_1(t) = \delta(t)$ and $x_0(t) = \delta(t - T)$. This pulse-sequence was repeated every 2000 time steps. After 400000 time steps the weight ρ was measured and plotted against the temporal difference T. The learning rate was set to $\mu = 0.001$. (d) Schematic explanation of the mutual weight change at a strong (A) and a weak synapse (B) with two subsequent delta pulses at the inputs x_1 and x_0 (for further explanations see text).

The corresponding weight change curves are plotted in Fig. 2.2b,c. The curves show that synaptic weights are strengthened if the presynaptic signal arrives *before* the postsynaptic signal and vice versa. The biological relevance of the learning curves becomes especially clear in the case $H_0 = H_1$. This learning curve with identical resonators is similar to the curves obtained in neuro-physiological experiments exploring spike timing dependent synaptic plasticity (STDP or "temporal Hebb") (Markram et al., 1997; Bi and Poo, 1998; Zhang et al., 1998; Abbott and Nelson, 2000). Furthermore in this case (Fig. 2.2b) it is seen that the location of

the maximum of the learning curve T_{opt} falls in the interval:

$$\frac{\lambda}{2\pi} < T_{opt} < \frac{\lambda}{4}, \qquad \frac{1}{2} < Q < \infty \tag{2.20}$$

where $\lambda = 1/f$ is the wave-length of the resonator.

The isotropic setup of the algorithm in principle also leads to weight changes at ρ_0. It is, however, evident that the change in ρ_0 is (very) small when the contribution from the other inputs ρ_k, $k \geq 1$ is small. This is most easily seen when considering Fig. 2.2d which shows a situation which arises after some learning by using the standard initial conditions. The size of the synapses depicts the momentarily existing weight values. The input sequence is such that a weight increase arises at synapse B from the influence of input line A onto line B ($+T$ in learning curve), whereas weight decrease occurs at synapse A because of the inverse causal ($-T$) influence of input line B onto line A. The degree of change is depicted by the plus and minus signs, showing that the decrease of A is smaller than the increase of B. For two similar inputs a simple rule of thumb is that the weight-change $\Delta\rho$ roughly follows the weight value of *the other* input scaled by the learning rate μ, while the sign of the change depends on the temporal sequence of events:

$$\Delta\rho_{late\ input} \approx \mu\ \rho_{early\ input} \tag{2.21}$$

$$\Delta\rho_{early\ input} \approx -\mu\ \rho_{late\ input} \tag{2.22}$$

As a result the strong input roughly maintains its strength while the contributions from the other inputs are small. This is the typical case when learning is guided by a strong reflex and the organism has the task to build up predictive pathways which should be weaker but more precise to prevent the disturbance.

The above obtained analytical results can be extended to cover the most general system structure as represented in Fig. 2.1 with $N > 1$. Equation 2.1 turns into:

$$V(s) = \sum_{k=0}^{N} \rho_k U_k(s) \tag{2.23}$$

keeping it in the Laplace-domain, because then it can directly obtained:

$$\Delta\rho_j(T) = \mu \frac{1}{2\pi} \int_{-\infty}^{+\infty} -i\omega V(-i\omega) U_j(i\omega) d\omega, \tag{2.24}$$

which is the general form of Eq. 2.12 in the LAPLACE domain. It should be noted

that for all $\Delta\rho_j$ this integral can still be evaluated analytically in the same way as in the special case with two resonators discussed above. In the following equations the index j is used for the input weights and k is used for the summation of the output-signal v.

2.4.2 Weight change when x_0 becomes zero

This section focuses on the weight development when the reference input (reflex) becomes silent ($x_0 = 0$) at some point during learning. This is motivated by the cases discussed in the introduction, where the goal of learning is to avoid (late, painful, damaging) reflex reactions. Thus, setting $x_0 = 0$ corresponds to the condition when the reflex has successfully been avoided. Note, that the circuit is left with just one (active) input (x_1) asking if its synaptic weight ρ_1 will continue to change.

The same restrictions (i–ii, p. 23) as above are used. Starting with equation 2.24 equation 2.23 is inserted into it. $x_0 = 0 \leftrightarrow X_0 = 0$ is set and the weight change becomes:

$$\Delta\rho_j = \mu\frac{1}{2\pi}\sum_{k=1}^{N}\rho_k\int_{-\infty}^{+\infty} -i\omega H_k(-i\omega)H_j(i\omega)d\omega \tag{2.25}$$

For $N = 1$ this results to:

$$\begin{aligned}\Delta\rho_1 &= \mu\frac{1}{2\pi}\rho_1\int_{-\infty}^{+\infty} -i\omega H_1(-i\omega)H_1(i\omega)d\omega &(2.26)\\ &= -\mu\frac{i}{2\pi}\rho_1\int_{-\infty}^{+\infty}\omega|H_1(i\omega)|^2 d\omega &(2.27)\end{aligned}$$

$H_1(i\omega)H_1(-i\omega) = |H(i\omega)|^2$ is valid since transfer functions can always expressed as products of complex conjugate pole-pairs. Multiplying $H_1(i\omega)$ with $H_1(-i\omega)$ leads to products of a complex number with its conjugate counterpart which renders the absolute value squared.

Since all transfer functions are symmetrical in relation to the real axis the frequency response $|H(i\omega)|^2$ is also symmetrical which leads to symmetrical responses in Eq. 2.27 at $|H_1(i\omega)|^2$. Due to ω in Eq. 2.27 the entire integral becomes anti-symmetrical and thus zero[1]. Thus, the weight ρ_1 stabilises if only x_1 is active.

[1] In a practical application (e.g., digital IIR filter) this is only true if the frequency responses

This result can be summarised in a rather intuitive way: With $N = 1$ and $x_0 = 0$ there is an input signal only at x_1. The weight change in that case is a correlation of a damped sine wave with its derivative which is a damped cosine wave. The correlation of a sine with a cosine is always zero.

In this thesis there will be no attempt to calculate the behaviour of the weights for $N > 1$, which is very tedious if not impossible. Instead simulation results will be shown for this later. However, the above argument can be extended by the Fourier theorem of wave decomposition to more inputs, because each sine wave from a resonator is multiplied by its cosine counterpart. Thus, also for $N > 1$ zero correlation is expected and a stop of the weight development as soon as there is no input (x_0).

2.5 Simulations

This section will perform simulations with the neuronal circuit of Fig. 2.1. The simulations have the purpose to validate the theoretical results from the last section and to explore the more complex situations (especially $N > 1$) which are not analytically tractable.

Simulations were performed under Linux using C++. Resonators were implemented as time-discrete IIR filters in the z-domain. The impulse-invariant transformation from the s-plane to the z-plane was used and the coefficients for the filters were calculated according to McGillem and Cooper (1984). Normalised time-steps were employed which result in normalised filter-frequencies in the range $f = [0 \ldots 0.5]$. In all applications frequencies less or equal to $f_{\max} = 0.1$ were used to avoid sampling-artifacts.

2.5.1 One filter in the predictive pathway: N=1

As before, the simplest case $N = 1$ is explored: one resonator in the reflex pathway x_0 and one resonator in the predictive pathway x_1 using the same restrictions as above (i-ii on page 23).

---

of the input X_1 and the transfer function H_1 vanish for high frequencies to avoid that the integral becomes ill defined $(\infty - \infty)$. In other words: the transfer functions must contain a *low-pass* term. This reflects the aspect that the time course of the input functions must be predictable (KALMAN filter-model, see Kalman 1960).

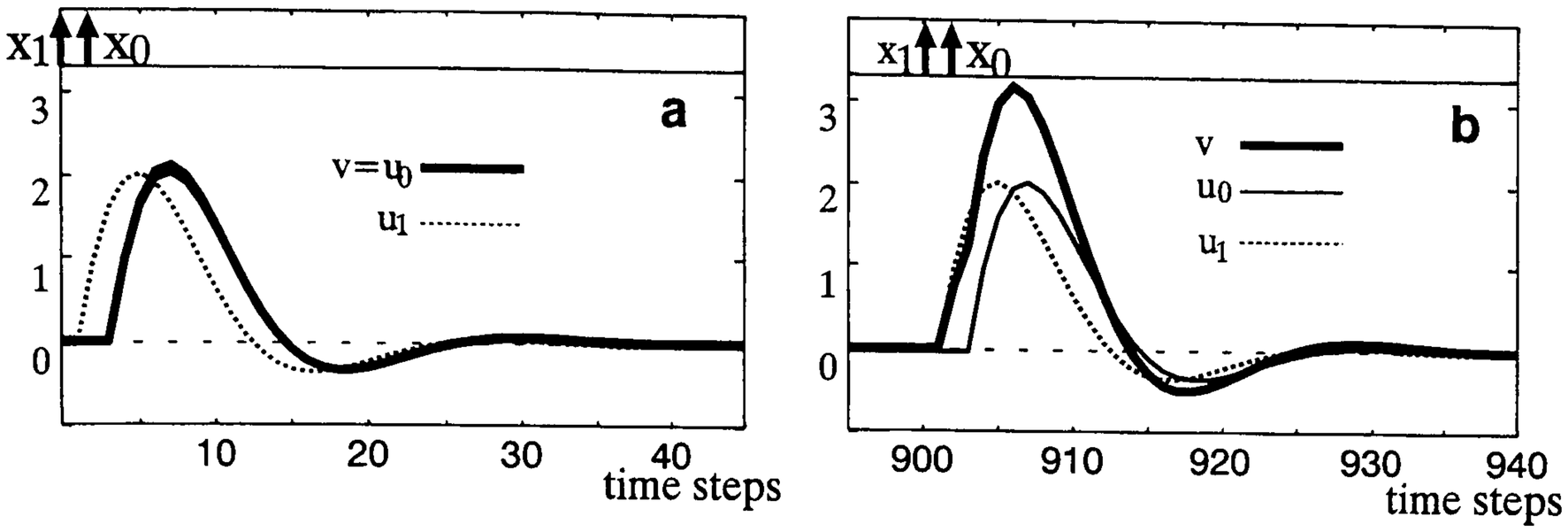

FIGURE 2.3: Simulation results with a circuit with two inputs, hence $N = 1$ (see Fig. 2.1). Input pulse sequences were repeated every 100 time-steps, the first starting at zero. Both resonators had values of $Q_{0,1} = 1$ and $f_{0,1} = 0.1$. The other parameters were $\mu = 0.01$ and $T = 2$. a) Result for time step 0, b) for time step 900.

2.5.1.1 Signal shape

Fig. 2.3a shows for time step 0 the δ-pulses at $x_{0,1}$ and the responses u_0 and u_1 from the resonators H_0 and H_1, respectively. Before learning the output v is identical to the signal u_0 because the weights were set to $\rho_0 = 1$ and $\rho_1 = 0$. The actual weight change of ρ_1 is caused by repeated pairing of the δ-pulses at x_0 and x_1. The result after 9 pairings is depicted in Fig. 2.3b. The comparison between Fig. 2.3a and Fig. 2.3b shows that the onset of the output v has shifted towards the earlier event x_1. Before learning it was identical to the resonator response u_0 in the reflex pathway. After learning the output is a superposition of both signals $u_{0,1}$ which leads to an onset which occurs together with the early onset of u_1. Thus, the circuit is able to "detect" the δ-pulse at x_1 as a predictor of the δ-pulse x_0.

2.5.1.2 Learning curve

Using the same setup the interval T can be varied. The change of ρ_1 as a function of T for the initial learning step (i.e., for $t = 0$ after one correlation) is considered. This was simulated using identical resonators $H_0 = H_1$ but also with different resonators $H_0 \neq H_1$. The results are shown together with the analytical findings in Fig. 2.2b,c having used the same parameters in both the simulation and the analytical calculation. Thus, the analytically calculated weight change curves are reproduced by the simulation results.

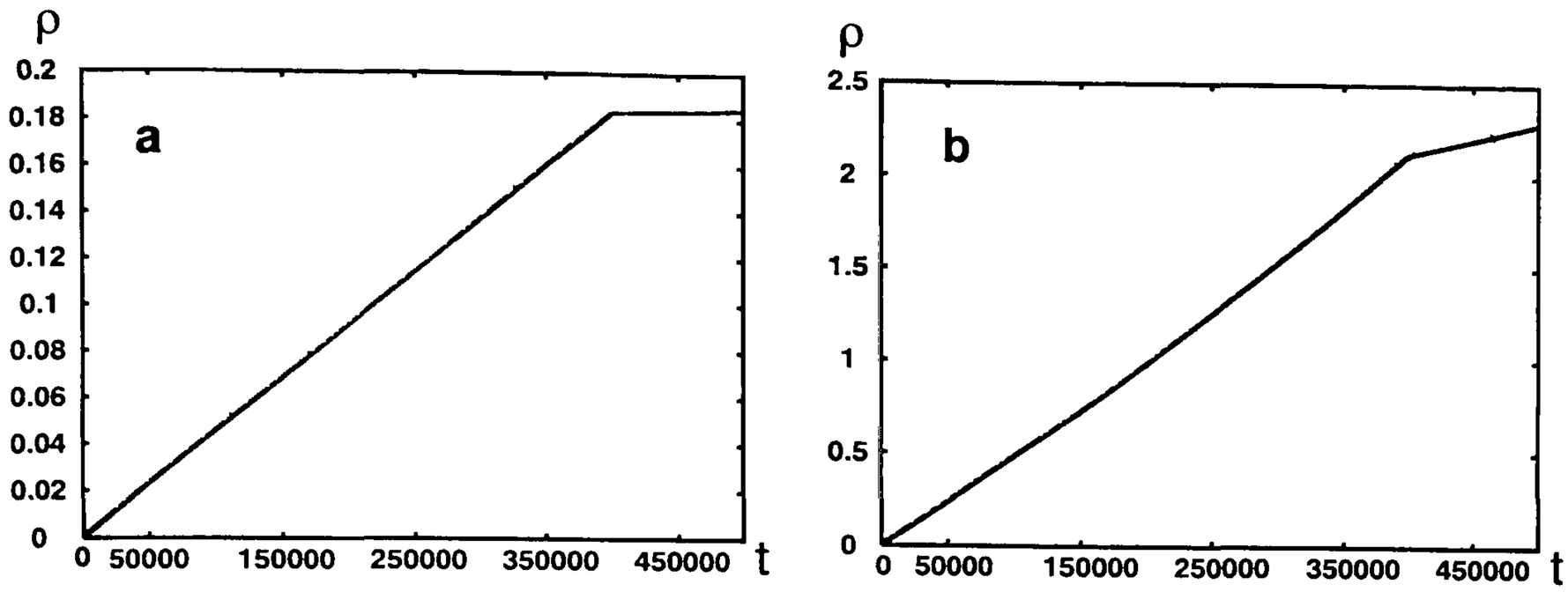

FIGURE 2.4: Simulated development of the weight ρ_1 for the case of two inputs ($N = 1$). Parameters were $f_{0,1} = 0.01$ and $Q_{0,1} = 1$. The inputs are triggered at a temporal difference of $T = 15$: $x_0 = \delta(t - T)$ and $x_1 = \delta(t)$. The pairing of the delta pulses is repeated every 2000 time steps. The learning rate is set to $\mu = 0.001$ in (a) and to $\mu = 0.01$ in (b).

2.5.1.3 Weight stabilisation for $x_0 = 0$:

The analytical results (Eq. 2.26) predict that ρ_1 should stabilise as soon as $x_0 = 0$. This, however, also requires that the learning rate μ is zero, which in reality cannot be ultimately achieved. The following simulation results show the effect of the learning rate on the development of the weights and compare the analytically obtained result with those obtained for more realistic situations. The simulation to test this was performed the following way: first the two resonators are triggered with paired δ-pulses. Then the input x_0 was switched off (i.e.: $x_0 = 0$) at $t = 400,000$ and only the input x_1 was still active.

Fig. 2.4 shows the weight development of ρ_1 over time for two different learning rates μ. With a low learning rate the weight ρ_1 approximately becomes constant when the input x_0 is switched off (see Fig 2.4a) whereas with a higher learning rate the weight continues to grow. With learning-rates too high the weight change during *one* correlation of two damped resonator responses must be taken into account in the correlation *itself*. Therefore, for example, Eq. 2.27 becomes a differential equation of ρ_1 which predicts an exponential growth of ρ_1. Therefore the learning-rate has to be adjusted in a way that the change of the weight during one correlation of two damped sine waves can be neglected.

Weight stabilisation is very desirable during learning (when the "desired state has been reached") but so is a high learning rate. These conflicting demands therefore lead to a trade-off, which needs to be taken care of in practical applications and

the right learning rate can be determined by the simulation shown here.

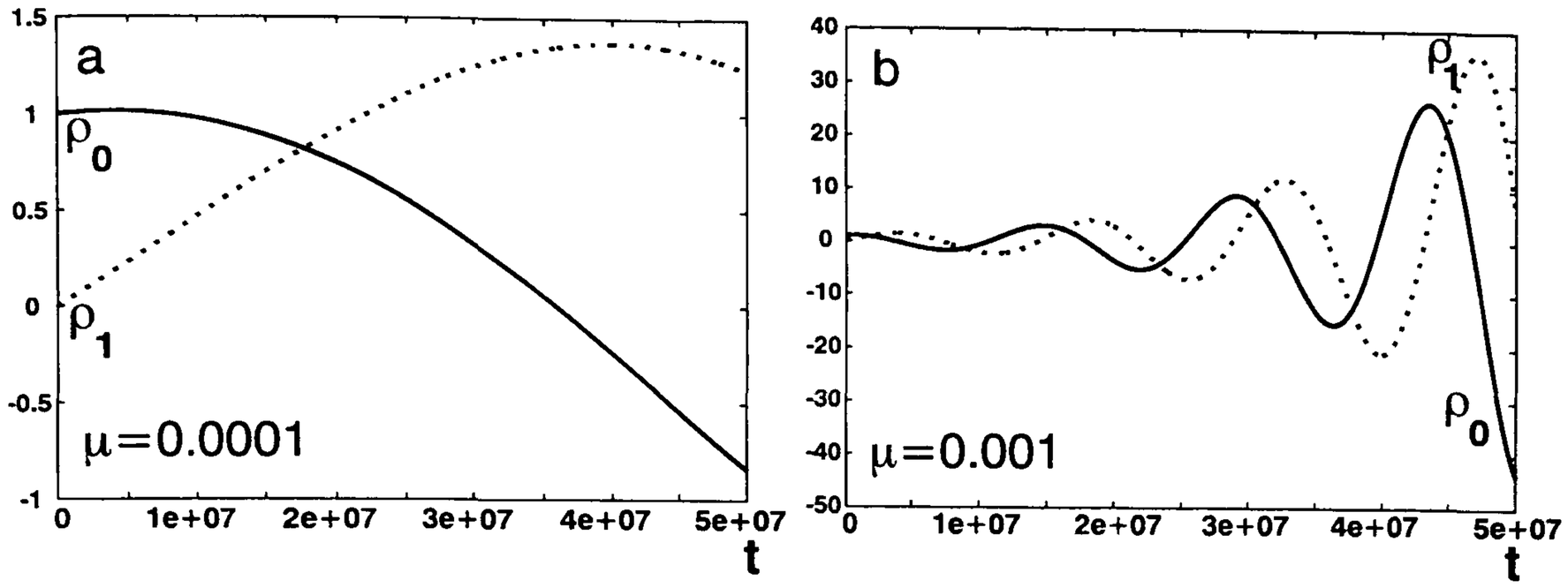

FIGURE 2.5: Simulated development of the weights ρ_0 and ρ_1 for the case of two inputs ($N = 1$). Parameters were $f_{0,1} = 0.01$ and $Q_{0,1} = 1$. The inputs are triggered at a temporal difference of $T = 15$: $x_0 = \delta(t - T)$ and $x_1 = \delta(t)$. The pairing of the delta pulses is repeated every 2000 time steps. The learning rate is set to $\mu = 0.0001$ in (a) and to $\mu = 0.001$ in (b).

2.5.1.4 Development of ρ_0:

In all cases discussed so far both weights were allowed to change, while it has been claimed ρ_0 remains stable. An easy intuition why this basically holds can be gained by using the "rule of thumb" defined above (Eq. 2.21,2.22). From this it is clear that the change of ρ_0 remains tiny for a prolonged time in the setup because ρ_1 equals zero at the beginning and μ is very small. Fig 2.5 shows the results from very long simulations with variable ρ_0. With a low learning rate (a) it can be seen that ρ_0 starts to change by more than 1% only after about 50000 learning steps (i.e. 25 pairings, and $\rho_1 = 0$, $\rho_0 = 1$ as the usual initial conditions). Even after 10^7 learning steps (i.e. 5000 pairings) the change of ρ_0 still can be neglected whereas ρ_1 has changed from zero to 0.5.

However, in some cases a strong decrease to $\rho_0 = 0$ is acceptable or even desirable if ρ_0 is the reflex-input and is no longer needed. Weight change should stop in this case at the moment when $\rho_0 = 0$. It makes sense to force the weight ρ_0 to zero after the condition $\rho_0 = 0$ has been reached by learning. A reversal of the sign of ρ_0 is not the desired behaviour since it would make the reflex via ρ_0 senseless. Therefore a more mild condition is to prevent ρ_0 from changing its sign. This would give ρ_0 the chance to grow again if the timing at the inputs is reversed and the reflex is needed again. From the moment ρ_0 has arrived at zero and kept at zero the output is only driven by the input x_1 via ρ_1. The paragraph which tested

the condition $x_0 = 0$ provides the answer how the weight ρ_1 behaves in the case $\rho_0 = 0$. From Eq. 2.27 it has been concluded that the weight ρ_1 does not change if the only contribution to the output v comes via ρ_1. Therefore the weight ρ_1 stabilises if ρ_0 is kept at zero or is not allowed to change its signs.

With a higher learning rate μ the system begins to oscillate and the weights are no longer stable (b). This oscillation can also be explained by the findings from the paragraph where the condition $x_0 = 0$ has been tested. In this paragraph only ρ_1 was allowed to change and the high learning rate lead to a differential equation of first order of ρ_1 (see Eq. 2.27). Here, both ρ_0 and ρ_1 are variable which leads to two coupled differential equations (see Eq. 2.24). Due to the coupling of the first-order differential equations means that they are able to generate oscillatory behaviour. However, this is not a desirable feature as the weights grow endlessly. As in the case above the learning rate has to be chosen in such a way that oscillation does not occur or that the wavelength of the oscillation is longer than the lifetime of the organism.

If there are more than two inputs $N > 1$ then the condition arises that after ρ_0 has been eliminated the other weights are freely floating and they have approximately the same strength. In Fig. 2.5 this is the case when $\rho_0 = \rho_1$ after 1/3 of the time course. Since ISO learning does not rely on the past the moment $\rho_0 = \rho_1$ can be taken as a starting point. Having two equal strong weights leads to a competition between them where the weights associated with early signals grow and the weights with later signals will get weaker. This leads at the end to the situation that the earliest signal will have the strongest weight and the latest signal will have the weakest or the weight will get the opposite sign.

Summarising, it can be seen that the reflex pathway will stay strong for a long time during learning so that the other weights have a chance to grow. This guarantees that during learning the reflex pathway is still functioning and only later it will be eliminated. However, this elimination would only happen if the reflex pathway would still be triggered. In the condition of "reflex avoidance" the reflex pathway will never be triggered again and therefore learning stops although there is a non-zero weight. This, on the other hand, always guarantees a fall-back to the reflex as a last resort.

Furthermore, in conditions where it is life threatening to unlearn the reflex it is obviously advantageous to force the weight ρ_0 to a fixed value to insure that the reflex can always be used in an emergency.

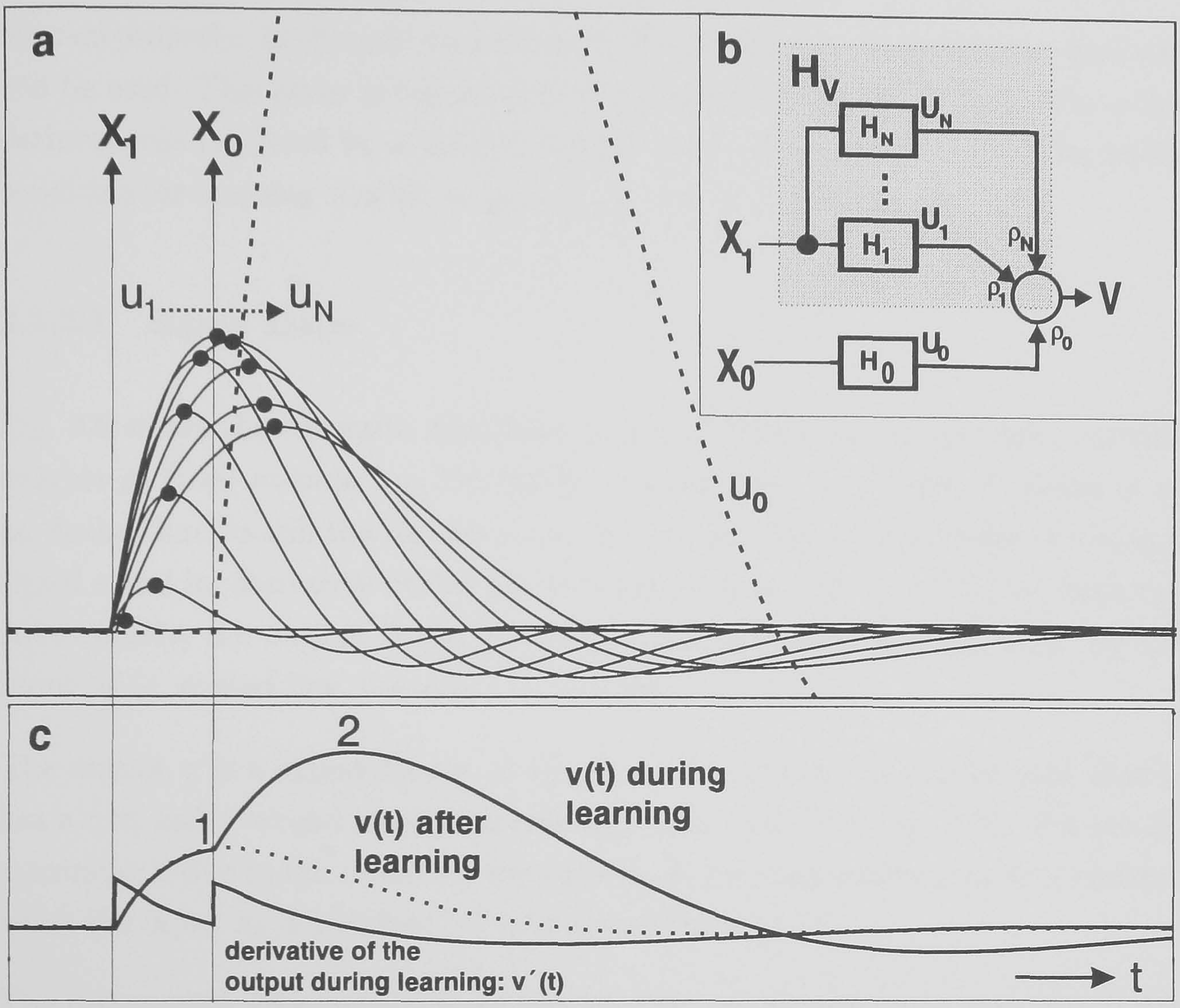

FIGURE 2.6: Multiple filters ($N = 10$) in the predictive pathway: Filter responses (a), the neuronal circuit (b) and its output during learning and after learning (c). The neuronal circuit (b) consists of a filter bank where the filter frequencies are set to $f_k = \frac{5f_0}{k}$; $k \geq 1$ and $f_0 = 0.01$. The learning rate was set to $\mu = 0.0005$ and $Q = 1$. The filter bank gets two different inputs $x_0(t) = \delta(t)$ (reflex-pathway) and $x_1(t) = \delta(t - T)$ (predictive pathway), $T = 10$. The delta pulses are repeated every 2000 time steps. After the 400,000 time steps x_0 is set to zero. The contribution of the signals $u_k\rho_k$ to the output v triggered by $x_1(t)$ is called H_V and is marked by the shaded box in (b). The weighted resonator responses $\rho_k u_k$ after learning are shown in (a). The output signals during learning (time step 390000) and after learning (after time step 400000) are shown in (c).

2.5.2 More than one filter in the predictive pathway

The setup with only one resonator ($N = 1$) in the predictive pathway has the disadvantage that there is only one specific temporal interval T_{opt} where learning (weight change) has the maximal rate. The use of an array of resonators with different frequencies in the predictive pathway removes this disadvantage (see inset in Fig. 2.6). The system should now be able to learn more than only one time

interval properly. In this section an array of 10 resonators in the predictive pathway will be used. This array is triggered with the same δ-pulse ($x_1 = \delta(t)$). The reflex pathway was triggered by a delayed δ-pulse ($x_0 = \delta(t - T);\ T = 10$). The initial condition for learning was set to $\rho_0 = 1;\ \rho_k = 0;\ k \geq 1$ as before.

2.5.2.1 Signal shape

Fig. 2.6 shows the resonator responses u_k scaled with their momentarily existing weights ρ_k (top) at time $t = 390,000$ during learning. The scaled response of u_0 (a, dashed line) is still the biggest at this time. The diagram also shows the output signal v and its derivative during the learning process (also $t = 390,000$, bottom). Additionally, the output signal is shown which is generated when silencing the input x_0 (c, dotted line, bottom, $t = 400,000$).

The output v is a superposition of all resonator outputs. It can be seen that it has a first and a second maximum (marked with 1 and 2 in Fig. 2.6). The second maximum is due to the resonator response from the reflex pathway u_0 and vanishes when the input x_0 is switched off (see dotted curve in c).

The first maximum is generated by superposition of the responses $\rho_k u_k, k > 0$ (i.e. all except u_0). In general this superposition process will always try to generate the first maximum as close as possible to x_0. This can be understood by the ongoing amplification of an initially existing asymmetry in the system in the following way. At the first learning step the derivative of v is zero before x_0 and then follows the shape of the v'-curve as shown in the diagram. Thus, there is one resonator response whose shape matches the v'-curve best (best positive correlation). Obviously, it is that particular resonator which has its maximum at (or closest to) the maximum of the v'-curve (second cusp, first is still zero). For this resonator the highest correlation result is obtained (Eq. 2.12) and, thus, the strongest weight-growth occurs at the beginning of learning. The other weights grow less strongly and their growth rate is approximately (inversely) related to the distance of their resonator maximum from x_0. This results in a distribution of weight values which follows the shape outlined by the y-position of the resonator maxima as shown in the top panel by the dots on the curves. Thus, superposition of these weighted responses leads to a maximum of v at x_0. This line of argumentation continues to hold also for the following learning steps, because the theoretical results suggest that the contribution of the correlation of the first part of the v'-curve (first cusp) with the $u_k, k > 0$, which would correspond to homo-synaptic learning, is zero in

all cases (see Eq. 2.25-2.27) thereby not affecting the weight change. Thus weight change continues to follow the distribution of the maximum in Fig. 2.6a. The resonator with the lowest frequency (f_l) determines the longest delay $T_{max} = \frac{1}{f_l}$ which can be learned. Equivalently the shortest delay is $T_{min} = \frac{1}{f_h}$ where f_h is the resonator with the highest frequency. Within the range $[T_{min}, T_{max}]$ any T causes an output with a maximum which always coincides with the location of x_0, provided there are enough resonators to allow for a sufficiently accurate superposition process.

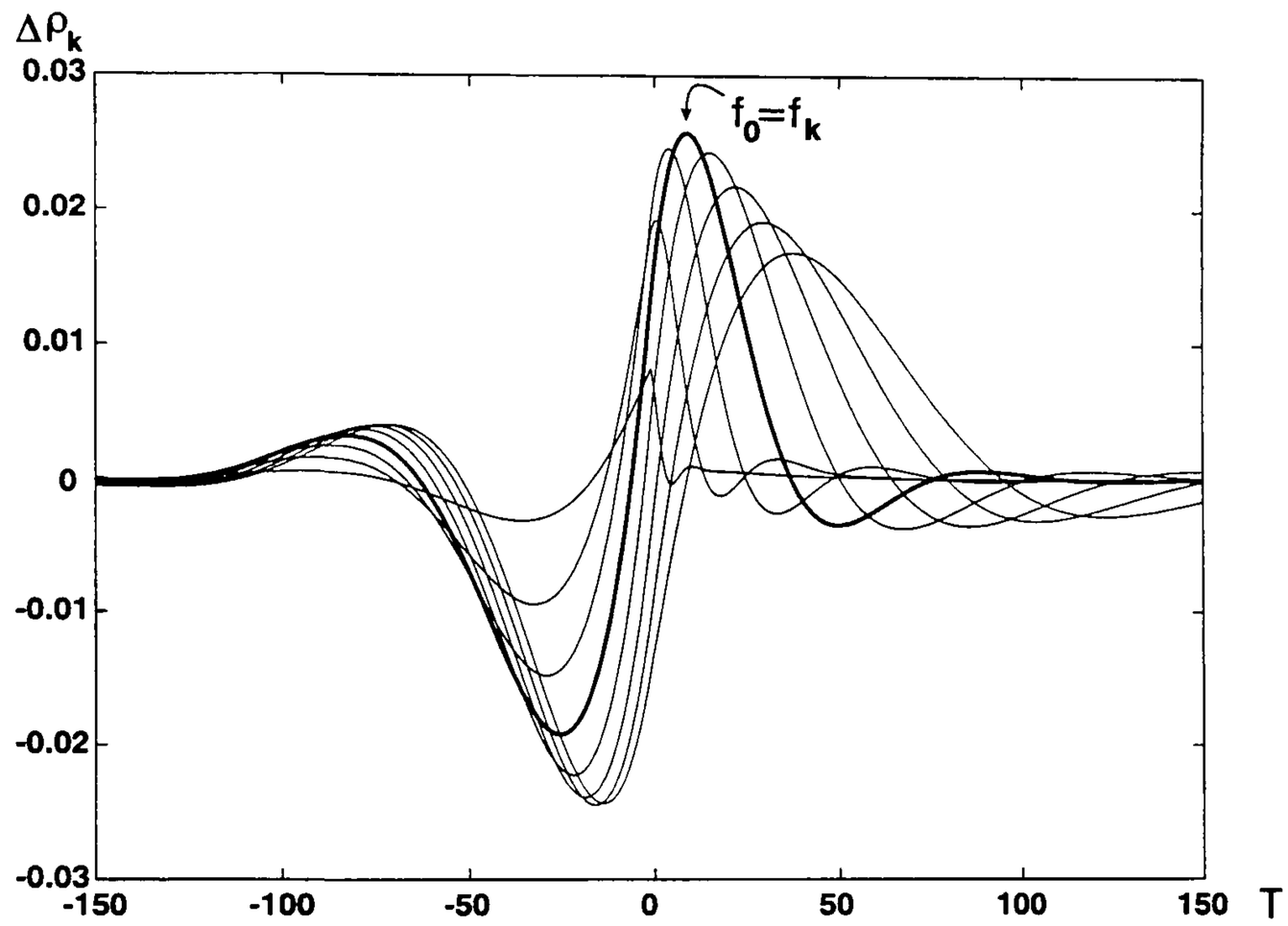

FIGURE 2.7: Weight changes ρ_j dependent of the temporal distance T with a filter bank of resonators ($N = 15$) set up as in Fig. 2.6b. The filter frequencies are set to $f_k = \frac{5f_0}{k}$; $k \geq 1$ with $f_0 = 0.01$ and $Q = 1$. The learning rate was set to $\mu = 0.0001$ and $Q = 1$. The case $f_0 = f_k$ is marked with a thick line and reproduces the curve in Fig.2.2b. The filter bank gets two different inputs $x_1(t) = \delta(t)$ (predictive pathway) and $x_0(t) = \delta(t - T)$ (reflex pathway). The delta pulses are repeated every 2000 time steps. After the 400,000th time step the weight ρ_j was measured and plotted against to the temporal difference T. Only every second curve is plotted.

2.5.2.2 Learning curve

As in the case of only two resonators; the dependence of the weight change on the temporal distance T can be explored. Now, however, there have to be monitored N changeable weights. For this experiment the same standard setup has been chosen using paired δ-pulses with a temporal delay of T, but now with 15 resonators ($N = 15$) in the predictive pathway. Their frequencies are chosen such that 10

resonators have a frequency which is higher and 5 resonators one which is lower than f_0 (see Fig. 2.7). Every second weight change curve is shown in Fig. 2.7 for $t = 0$ where T was varied from -150 to 150. Every curve in this diagram represents one weight ρ_k of a specific resonator h_k as a function of T. The curve plotted with the thick line belongs to the resonator h_k which has the same frequency as the resonator h_0, hence $f_k = f_0$. The other weight change curves belong to resonators in the predictive pathway which have different frequencies compared to f_0. It can be seen that every weight change curve has a specific T where weight change is maximal or (in support of the argument used to explain the first maximum in Fig. 2.6). Or the other way round: for specific values of T and large N there exists always one particular resonator which shows maximum weight change.

Another interesting result is that the weight change curve with $f_k = f_0$ is identical to the weight change curve with only one resonator (see Fig. 2.7). The fact that both weight change curves are the same is due to the linearity of ISO learning.

In summary, in an array of different resonators every resonator is only responsible for a specific and limited range of temporal intervals so that such an array is able to cover a wide range of different temporal intervals. The weight change curves for the different weights give precise information on which resonator yields the maximum contribution to the output signal.

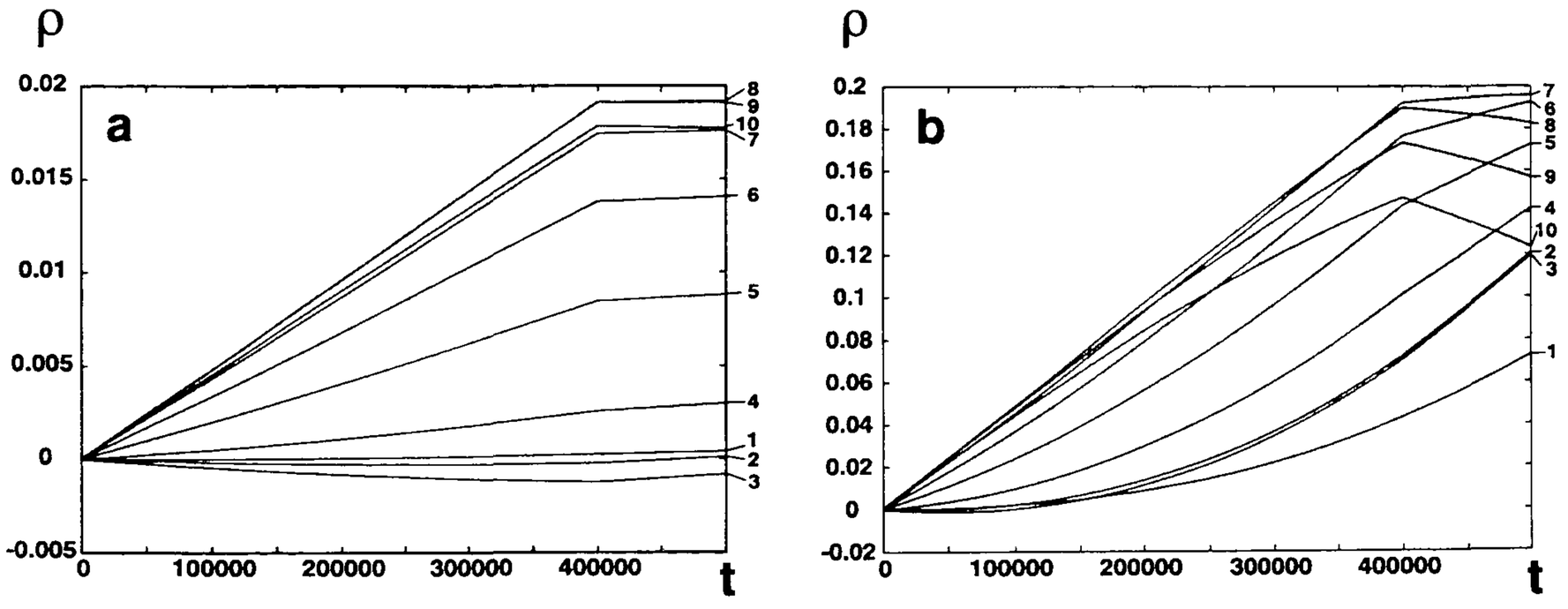

FIGURE 2.8: Weight change of multiple resonators $N = 10$ in dependence of the learning rate. The neuronal circuit (see Fig. 2.6b) consists of a filter bank where the filter frequencies are set to $f_k = \frac{0.1}{k}$; $k \geq 1$ where the index k is also used as a label for the different curves in this figure ($Q = 1$ in both cases). The filter bank gets two different inputs $x_1(t) = \delta(t)$ (predictive pathway) and $x_0(t) = \delta(t - T)$ (reflex pathway) with $T = 10$. The delta pulses are repeated every 2000th time step. After 400,000 time steps x_0 was set to zero. The learning rate was set to $\mu = 0.0001$ in (a) and to $\mu = 0.001$ in (b).

2.5.2.3 Weight stabilisation for $x_0 = 0$:

The next question that arises is if the weights also stabilise in a *multi*-resonator setup if the reflex pathway x_0 becomes zero (see Fig. 2.8). The same setup as before was used for the simulation ($N = 10$ and paired δ-pulses with $T = 10$). The test was performed in the same way as above by setting x_0 to zero at time $t = 400,000$. Fig. 2.8 shows that the weights stabilise in the limit of $\mu \to 0$. Thus, again the crucial parameter for an approximate weight stabilisation is the learning rate μ, which is too high in Fig. 2.8b.

Because of the complexity of the mathematics in a setup with filter-banks, it is not possible to give robust analytical arguments for weight stabilisation in the multi-resonator case. However, the argument from the case with one resonator ($N = 1$) can be used here, namely that the individual resonator responses (sine-waves) are orthogonal to the derivative of the output (cosine wave) as soon as $x_0 = 0$, (see dashed curve in Fig. 2.7) leading to zero value of the correlation integral. The experimental findings in Fig. 2.8 support this notion. Thus, also in the multi-resonator case the desired property of weight stabilisation for $x_0 = 0$ is obtained in the limit of $\mu \to 0$.

2.6 Summary

In this chapter the internal structure of the organism has been presented. First, the relation between the sensor inputs and the motor output was introduced: in a first processing stage all sensor inputs are bandpass filtered. In a second stage these bandpass filtered signals create a weighted sum which directly represents the motor output. All inputs are treated equally. However the input which is associated with the reflex should have initially a strong weight.

Learning takes place according to Eq. 2.3. A weight is changed by correlating the corresponding filtered sensor signal by the derivative of the motor output. This learning guarantees that sequence learning takes place. Sensor inputs which precede the output signal will strengthen their corresponding weights and sensor inputs which lag behind will weaken their corresponding weights.

Every bandpass is tuned to a specific temporal delay. This is a disadvantage in situations where the temporal delay is not known a priori. To learn unknown temporal delays different resonators have to be combined. This leads to the ap-

plication of filter-banks. One sensor signal is fed into a filter bank with different frequencies so that every filter covers a certain temporal delay.

Weight stabilisation is an important property as it marks the success of learning. It has been proven that if only one input is triggered all weights stabilise. Simulations have shown that this analytical finding can be generalised: weight-stabilisation is also possible if more than one input is triggered, for example in a filter-bank. The weights stabilise if all active inputs are triggered synchronously.

Thus, it has been shown that it is possible to establish an organism and a learning rule which meets all the requirements introduced at the beginning of this chapter.

Chapter 3

The Organism in its Environment

3.1 Introduction

In the preceding chapter only the organism has been described while its environment has been ignored. In this chapter the environment will be introduced which provides the feedback from the organism's motor output to its sensors. Thus, a closed loop will be formed. As in the previous chapter, the aim is to arrive at a mathematical description of the closed loop condition, obtain analytical results and support them by simulations.

As pointed out in section 1.7 the simplest closed loop control is reactive control. It is robust and needs only limited information about the environment. However, reactive control has a disadvantage in that it is always too late. The solution is pro-active control which anticipates the trigger of the reactive control loop. ISO learning seems to be a candidate which turns a reactive system into a proactive system. Therefore the central problem in this chapter is whether or not ISO learning is able to eliminate the disadvantage of reactive control, namely of always being too late.

The following sections will show that ISO learning is able to turn a reactive system into a proactive system. Consequently, the first section starts with a formal reactive system (section 3.2). On top of this reactive system ISO learning will be introduced (section 3.3). This enables the organism to overcome its reactive behaviour and replace it with proactive behaviour. This will be shown analytically and also by computer simulations.

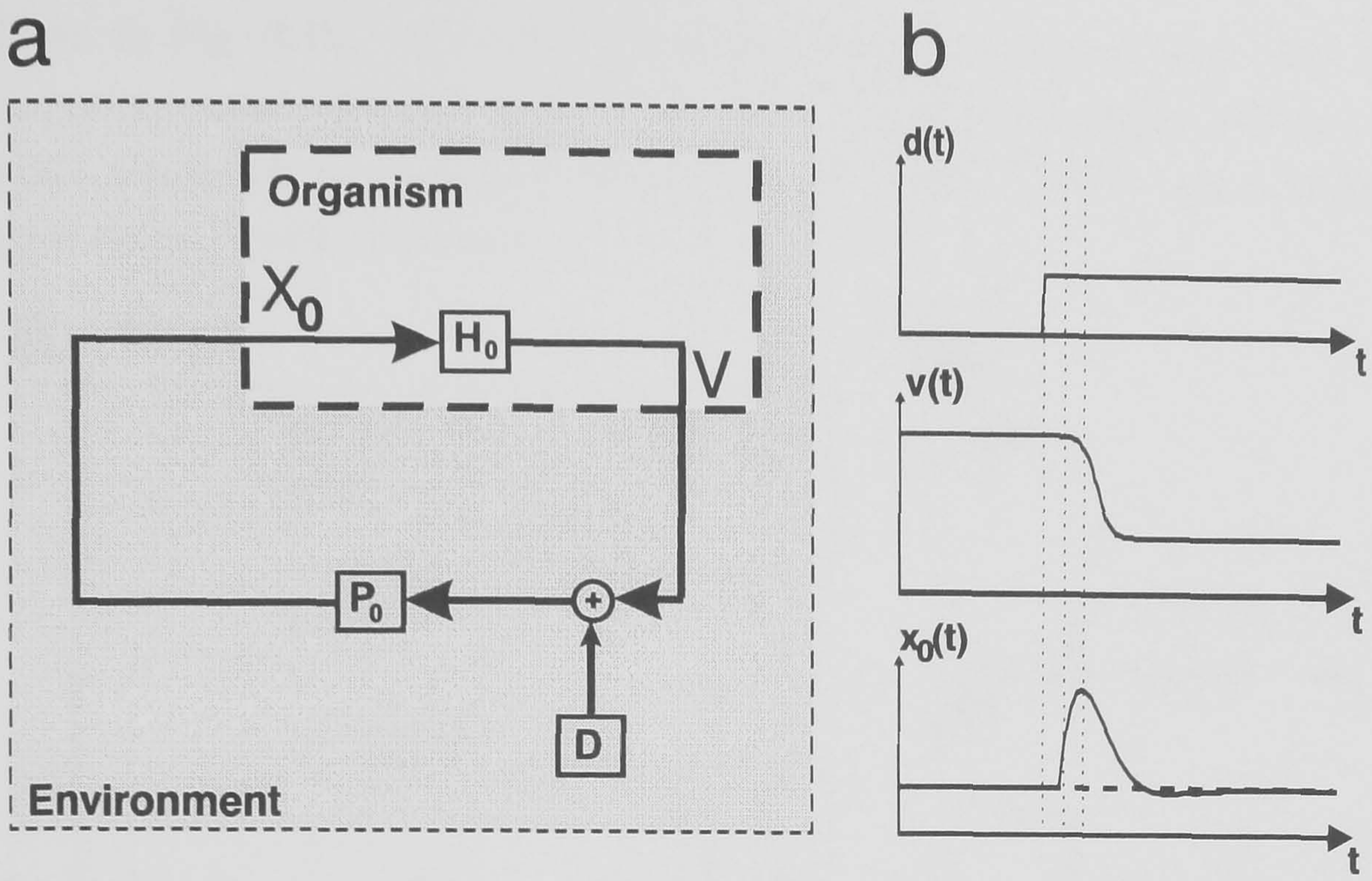

FIGURE 3.1: Fixed reflex loop: the organism transfers a sensor event X_0 into a motor response V with the help of the transfer function H_0. The environment turns the motor response V again into a sensor event X_0 with the help of the transfer function P_0. In the environment there exists the disturbance D which adds its signal at $\oplus$ to the reflex loop. **b)** 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.

3.2 Reflex loop behaviour

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 precisely and fast as possible. In the model presented in this thesis it is assumed that the desired state of the reflex feedback loop is unchanging and defined by the properties of the reflex loop, namely that the reflex has to be eliminated. Therefore it is defined as $X_0 = 0$ (e.g., "no collision should be felt"). First the system is discussed without learning. Fig. 3.1a shows the situation of a learner embedded into a very simple but generic (i.e., unspecified) environment which has a transfer function P_0. This learner is able to react to an input only by means of a reflex. Consider the case of obstacle avoidance. If an obstacle is encountered (disturbance D) and felt by collision-sensors (X_0) the unconditioned retraction reflex performs an avoidance reaction (scheduled by the transfer function H_0) trying to re-establish the desired state ($X_0 = 0$).

A possible set of signals (in the time-domain) which can occur in such a system

is shown in Fig. 3.1b. 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 to restore the desired state $x_0 = 0$. Thus, there is always a reaction-delay in such a system.

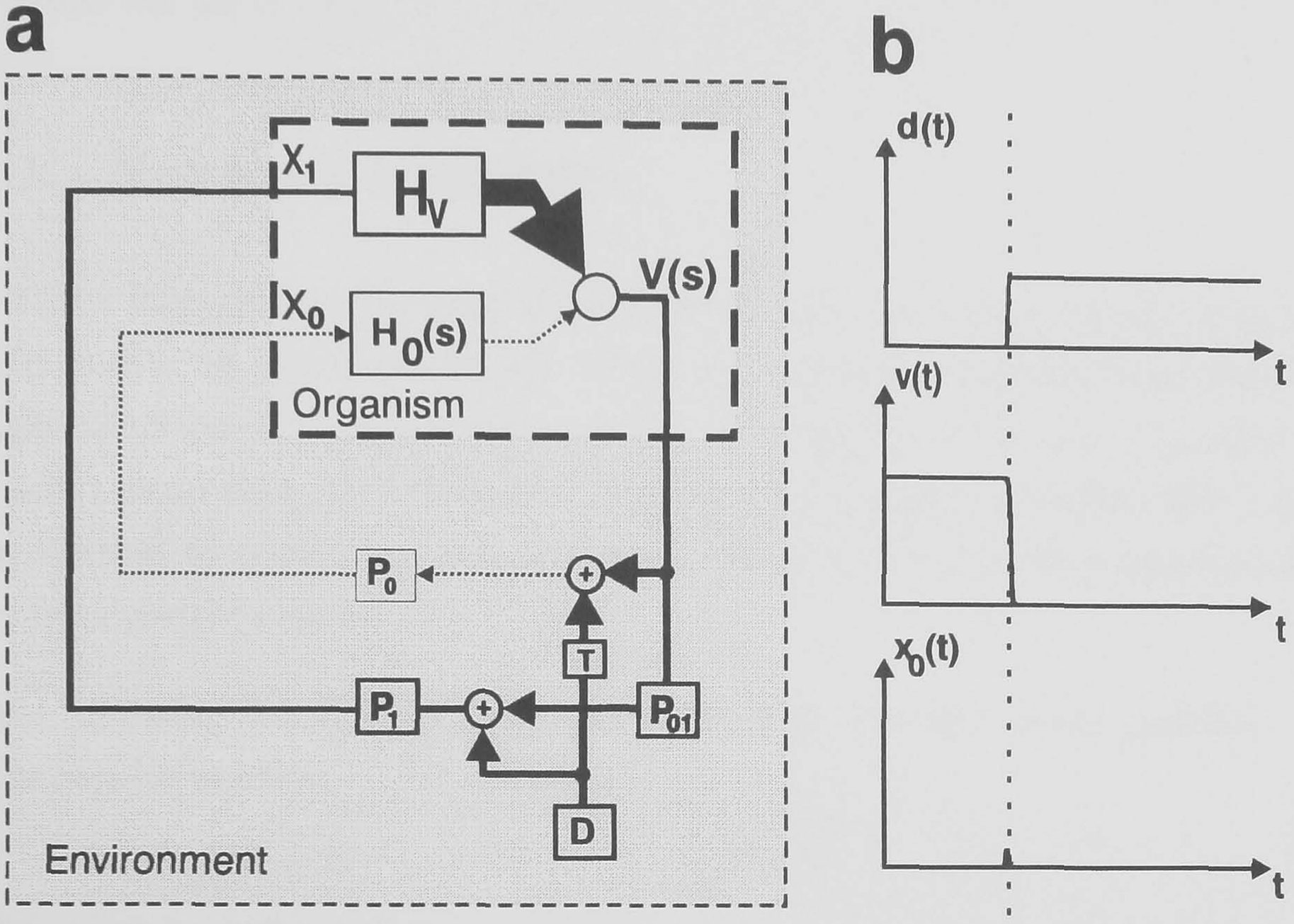

FIGURE 3.2: Schematic diagram of the augmented closed loop feedback mechanism which now contains a secondary loop representing temporal sequence learning. a) H_0 and P_0 form the inner feedback loop already shown in Fig. 3.1. The new aspect is the input-line S_1 which gets its signal via transfer function P_1 from the disturbance D. The inner feedback loop receives a delayed version (τ) 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 S_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.

3.3 Augmenting the reflex by temporal sequence learning

In this section it will be shown that the ISO learning algorithm can approximate the inverse controller of the reflex. Fig. 3.2 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 after a delay T (filtered by P_0). The goal of ISO 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 which is depicted by Fig. 2.6b. The environmental transfer function P_{01} closes the outer loop.

3.3.1 Necessary 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 it must be shown what shape the transfer function of the predictive pathway H_v (see Figs. 2.6b and 3.2) takes with the assumption that $X_0 = 0$ holds. This is the necessary condition, which needs to be obeyed to obtain an appropriate H_v. It generally applies *regardless* of the learning algorithm used.

In the following the function argument s will be omitted where possible. The inputs can be written:

$$X_0 = P_0[V + De^{-sT}] \tag{3.1}$$

as the reflex pathway and

$$X_1 = \frac{P_1 D + X_0 H_0 P_{01} P_1}{1 - P_1 P_{01} H_V} \tag{3.2}$$

$$H_V = \sum_{k=1}^{N} \rho_k H_k \tag{3.3}$$

as the predictive pathway (see Fig. 3.2). Eliminating X_1 and V results to:

$$X_0 = e^{-sT} D + H_V \frac{P_1 D + X_0 H_0 P_{01} P_1}{1 - P_1 P_{01} H_V} \tag{3.4}$$

Solving for $X_0 = 0$ leads to:

$$H_V = \sum_{k=1}^{N} \rho_k H_k \tag{3.5}$$

$$= -\frac{P_1^{-1} e^{-sT}}{1 - P_{01} e^{-sT}} \tag{3.6}$$

The transfer function H_V is the overall transfer function of the predictive pathway. Eq. 3.5 demands that the weights ρ_k should be adjusted in such a way that Eq. 3.6 is obtained at the end of learning.

Eq. 3.6 requires interpreting. First, the numerator is discussed, keeping in mind 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. 3.6 compensates the delay between the signal in X_1 and that at X_0, when the disturbance actually enters the inner feedback loop.

Now the relevance of the denominator has to be discussed showing that it can be generally ignored. Transfer functions are fully described by their poles and zeroes. Poles very strongly affect the behaviour of a system, while zeroes 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, Eq. 3.6 can be rewritten as:

$$H_V = -P_1^{-1} e^{-sT} \frac{1}{1 - P_{01} e^{-sT}} \tag{3.7}$$

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". It is, thus, an entity which violates causality.

As a result, there are no additional poles for H_V and in the following it is 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]

$$H_V = P_1^{-1} e^{-sT} \tag{3.8}$$

The last equation represents the necessary condition for the learning and the next

[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 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.

two sections will ask the question if ISO learning is sufficient to achieve this.

3.3.2 Solutions in the steady state case X_0=0

Here it is shown that for one resonator there already exists a solution which approximates Eq. 3.8 to the second order. Results for a forth order approximation have been numerically obtained, showing that the approximation continues to improve.

Thus, first the discussion is limited 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. It will be specified which parameters the resonator H_1 in the outer loop has to satisfy the learning goal. At first $P_1 = 1$ is set, looking at the case when the environment does not alter the shape of the disturbance (but see below).

Considering Eq. 3.8 and Eq. 3.3

$$-e^{-sT} = \rho_1 H_1 \tag{3.9}$$

The resonator H_1 has two parameters $f_1 = 1/T_1$ and Q_1 and together with its weight ρ_1 there are three parameters which solve this equation and have to be determined.

The left hand side of Eq. 3.9 can now be developed into a Taylor series:

$$-\frac{1}{e^{sT}} = \frac{-1}{1 + sT + \frac{1}{2}s^2T^2 + \ldots} \approx \frac{-2T^{-2}}{2T^{-2} + 2sT^{-1} + s^2} \tag{3.10}$$

and the right hand side of Eq. 3.9 has to be explicitly written out according to Eqs. 2.2–2.5:

$$\rho_1 H_1(s) = \frac{\rho_1}{(s+p)(s+p^*)} = \frac{\rho_1}{\underbrace{pp^*}_{(2\pi f_1)^2} + s\underbrace{(p+p^*)}_{\frac{-2\pi f}{Q_1}} + s^2} \tag{3.11}$$

The coefficients of Eq. 3.10 can now be compared with Eq. 3.11 and the resulting parameters are:

$$\rho_1 = -2\frac{1}{T^2}, \qquad f_1 = \pm\frac{1}{\pi\sqrt{2}}\frac{1}{T}, \qquad Q_1 = \sqrt{\frac{1}{2}} \tag{3.12}$$

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 the resonator-frequency f_1 can be interpreted in the context of the simulations done in section 2.5.2. Remember that $X_0 = 0$ was set and hence $V = X_1 H_1$. If a δ-pulse at X_1 is considered, the impulse response of the resonator $h_1(t)$ at the output is:

$$v(t) = \rho_1 \frac{1}{b_1} \sin(b_1 t) e^{-a_1 t} = \rho_1 T \sin\left(\frac{t}{T}\right) e^{-\frac{t}{T}} \tag{3.13}$$

This function has its maximum at $t^{(2)}_{max} = T\text{atan}(1)$. The notation $t^{(2)}$ refers to the second order approximation. One can assume that this is approximately equal to $t^{(2)}_{max} \approx T$ (see below). This, however, would be indicative of a response maximum which occurs at the moment where the input x_0 is to be expected. The reader is referred to section 2.5.2 where this type of behaviour has indeed been observed in the simulations (Fig. 2.6). In these simulations it has been found that during learning the output has always its first maximum at the location where x_0 occurs (or would have occurred). This shows that the experimentally observed convergence behaviour of the algorithms leads to a function H_v which has similar properties to that obtained from the second order Taylor approximation.

The relation $t^{(2)}_{max} \approx T$ could be confirmed because the same Taylor-approximation has been performed with $N = 2$ (leading to a fourth order Taylor approximation). The resulting set of equations has been solved numerically (with the commercial package "Derive") and the solution leads to $t^{(4)}_{max} = 0.978T$. This suggests that $t^{(\infty)}_{max} = T$ is correct in the limit of $N \to \infty$.

For all *practical* purposes N needs to be found in trying to resolve the tradeoff between the actually needed precision for $t^{(\infty)}_{max} \to T$ and hardware/software engineering constraints (costs). The robot experiment below will demonstrate that in a real world application already few resonators ($N = 10$) suffice to obtain the desired behaviour after learning.

Now more complex transfer functions for P_1 have to be considered. Up to this point P_1 has been set to 1 which means that the disturbance reaches the input X_1 un-altered which is in general not the case. Because of specific sensor-properties and 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). The behaviour of ISO learning with such complex input-functions can be derived if one recalls that a Taylor-approximation

of Eq. 3.9 has been used and matched with the sum of resonators to obtain the coefficients. This, however, allows us to conclude that any transfer function P_1 of the shape:

$$P_1 = \frac{(s+z_0)(s+z_0^*)\ldots(s+z_n)(s+z_n^*)}{(s+p_0)(s+p_0^*)\ldots(s+p_m)(s+p_m^*)} \tag{3.14}$$

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 rational function similar to that in Eq. 3.14 above[2]. 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, it can normally be assumed that the environment does not actively interfere with signal transmission in such a system and it can therefore – with great likelihood – be represented by Eq. 3.14.

A more intuitive explanation that the function P_1 does not change the overall behaviour of the learning circuit comes from the simulation results in section 2.5.2, especially Fig. 2.6. In this simulation of the multi resonator condition a maximum was achieved at the moment when the event x_0 was triggered or would have been triggered. This maximum was due to the strong derivative at the output when event x_0 occurs. Thus, the maximum will always be established as a there is a strong derivative and resonator-responses which can be correlated with this strong derivative.

Consider the case that $P_1 \neq 1$. In that case all resonators of the predictive pathway get a *filtered* version of the disturbance D: $X_1 = DP_1$. Consequently the resonator responses will differ from the case $P_1 = 1$. However, the learning rule will still correlate the resonator responses u_k with the output's derivative v'. As a consequence the resonator responses with the highest correlation with the derivative will give the strongest contribution to the output. Since the derivative is strongest at the moment x_0 is triggered the output still gets a maximum at the moment x_0 is triggered. Thus, it can be concluded that even with functions $P_1 \neq 1$ the output compensates the disturbance D and that the results generalise to more complex P_1.

Therefore, it can be argued that an appropriate approximation of the complete Eq. 3.8 will be found in almost all situations. The robot application which will be shown below supports this notion experimentally.

---

[2]Note that it is even possible to approximate zero crossings of Eq. 3.14 since it is a *sum* of resonator responses. If the overall transfer function of a sum of resonators $(H_1 + H_2 + \ldots)$ is calculated this leads automatically also to zero crossings which can be used to identify them with the zero crossings in Eq. 3.14. Thus, the approximation continues to hold including also phase terms.

3.3.3 Convergence Properties (sufficient condition)

The last section has shown that it is possible to construct approximate solutions of Eq. 3.8 using resonators so that $X_0(s) \to 0$. This section addresses the question 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 be found. An approach, which leads to success, however, is to apply perturbation theory instead.

The starting point of such an analysis is that a set of weights ρ_k, $k > 0$ has been found which solves Eq. 3.8. It is known that the development of the weights follows Eq.2.23. Now the system is perturbed by substituting ρ_j in Eq.2.23 with $\rho_j + \delta\rho_j = \tilde{\rho}_j$. To assure stability it must be proven that the perturbation is counteracted by the weight change. Thus Eq.2.23 must be solved hoping to find:

$$\Delta\rho_j \sim -\delta\rho_j \tag{3.15}$$

This must even hold for strong changes $\delta\rho_j$ so that convergence is guaranteed. Therefore any approximation (like a Taylor series in $\delta\rho_j$) is not permitted.

The signals U and V have to be defined. The signal U is easy as it is simply the filtered input X.

$$U_j = X_j H_j = \begin{cases} X_0 H_0 & \text{for} \quad j = 0 \\ X_1 H_j & \text{for} \quad j > 0 \end{cases} \tag{3.16}$$

V is more complicated. The definition (Eq. 2.1) provides:

$$V = \rho_0 X_0 H_0 + X_1 \sum_{k=1}^{N} \rho_k H_k \tag{3.17}$$

and from above it is known (Eq. 3.1):

$$X_0 = P_0[V + De^{-sT}] \tag{3.18}$$

Thus for V this results in:

$$V = \rho_0 P_0 [V + De^{-sT}] H_0 + X_1 \sum_{k=1}^{N} \rho_k H_k \tag{3.19}$$

$$= \rho_0 P_0 H_0 V + \rho_0 P_0 H_0 De^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k \tag{3.20}$$

yielding:

$$V = \frac{\rho_0 P_0 H_0 De^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k}{1 - \rho_0 P_0 H_0} \tag{3.21}$$

Substituting $\rho_j \rightarrow \rho_j + \delta\rho_j$ leads to:

$$\tilde{V} = \frac{\rho_0 P_0 H_0 De^{-sT} + X_1 \sum_{k=1}^{N} \rho_k H_k + X_1 \sum_{k=1}^{N} \delta\rho_k H_k}{1 - \rho_0 P_0 H_0} \tag{3.22}$$

$$= V + \frac{X_1 \sum_{k=1}^{N} \delta\rho_k H_k}{1 - \rho_0 P_0 H_0} \tag{3.23}$$

Then calculating the weight change using Eq. 2.23:

$$\Delta\tilde{\rho}_j = \frac{\mu}{2\pi} \int_{-\infty}^{\infty} -i\omega \left[V^- + \frac{X_1^- \sum_{k=1}^{N} \delta\rho_k H_k^-}{1 - \rho_0 P_0^- H_0^-} \right] X_1^+ H_j^+ d\omega \tag{3.24}$$

where the abbreviations $^+$ and $^-$ for the function arguments $+i\omega$ and $-i\omega$ have been introduced. The first part of this integral describes the equilibrium state condition and can be dropped, thus:

$$\Delta\rho_j = \frac{\mu}{2\pi} \sum_{k=1}^{N} \delta\rho_k \int_{-\infty}^{\infty} -i\omega \frac{|X_1|^2 H_k^-}{1 - \rho_0 P_0^- H_0^-} H_j^+ d\omega \tag{3.25}$$

where for X_1 it has been made use of the fact that for transfer functions in general it can be written: $X^+X^- = |X|^2$ where the superscripts $^+$ and $^-$ for the function arguments $+i\omega$ and $-i\omega$ has been used. This result is still general in the sense that Eq. 3.25 does not necessarily deal with resonator functions. So at this moment it is still possible to make some reasonable assumptions about the set of H_k. To avoid correlational effects between resonators with different parameters ($k \neq j$)

orthogonality is assumed, given by[3]:

$$0 = \int_{-\infty}^{\infty} -i\omega \frac{|X_1|^2 H_j^+ H_k^-}{1 - \rho_0 P_0^- H_0^-} d\omega \quad \text{for} \quad k \neq j \tag{3.26}$$

This condition can be used to simplify Eq. 3.25 which leads to:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} |X_1^+|^2 |H_j^+|^2 \frac{-i\omega}{1 - \rho_0 P_0^- H_0^-} d\omega \tag{3.27}$$

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. This argument shall be discussed more concretely.

By the use of Plancherel's theorem (Stewart, 1960) the integral in Eq. 3.27 is transferred into the time-domain:

$$\Delta\rho_j = \mu\delta\rho_j \int_0^{\infty} a_{x*h}(t) f'(t) dt \tag{3.28}$$

where $a_{x*h}(t)$ is the autocorrelation function of $x_1(t) * h_j(t)$ which is the inverse transform of $|X_1^+ H_j^+|^2$ ($*$ denotes a convolution). Note that the remaining term in Eq. 3.27: $\frac{-i\omega}{1-\rho_0 P_0^- H_0^-}$ contains the derivative operator $-i\omega$ in the numerator. Thus, $f'(t)$ in Eq. 3.28 is the temporal derivative of the impulse response of the inverse transform of $\frac{1}{1-\rho_0 P_0^- H_0^-}$.

At that point it must be asked 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. It can, however, in general be assumed that P_0 being an environmental transfer function should again behave passively and follow Eq. 3.14. Furthermore it is known 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 as a low pass smears out a sharp step response and therefore delays the transmission. As a consequence

[3]This orthogonality-assumption will be waived later and is used here to make the mathematics treatable. In the simulations later on it will be shown that the real resonators are not orthogonal to each other but the non-diagonal elements do not change the general result. Therefore the non diagonal elements are simply set to zero.

it can be stated that the fraction $\frac{1}{1-\rho_0 P_0 H_0}$ is dominated by the characteristic of a (non-standard) high-pass as the inverse of a low-pass becomes a 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 this especially holds with a delta-pulse as a disturbance at $t = 0$, corresponding to $x_1(t) = \delta(t)$.

Thus, for an orthogonal set of H_k, ISO learning will converge if P_0 is dominated by a low-pass characteristic and if the disturbance D has a short duration in relation to the reaction-time of the feedback loop.

Finally, it has to be proven, that Eq. 3.27 is zero in the equilibrium state case where the feedback loop is no longer needed. This leads to $0 = X_0 = \rho_0 H_0 P_0$ and the denominator becomes one. The weight change results in:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} -i\omega |X_1^+|^2 |H_j^+|^2 d\omega \tag{3.29}$$

This integral is anti-symmetrical and thus zero as required. In the open-loop condition there had been an equivalent result. There the synaptic weights stabilised as soon as explicitly $X_0 = 0$ was set (compare Eq. 2.27). 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.

3.3.4 Matching the theoretical convergence properties to the practical approach

3.3.4.1 Unity feedback loop

As stated above, in a real application, the reflex loop has to be stable. The above section simply *demanded* that the reflex loop has to be stable without explicitly specifying a reflex loop. An explicit definition has been avoided since the above derivations should be as general as possible. This section now introduces a specific feedback loop with real resonators. An analytical derivation is no longer possible but numerical simulations are performed with this concrete example. This example shall be kept as simple as possible without eliminating the important property of a reflex loop: the basic (critical) property is its delay characteristic. This property underlies the conceptual necessity for temporal sequence learning and it

was the essential property of the above mathematical treatments. The specific characteristics of some of the transfer functions, on the other hand, are secondary and can, therefore, be simplified.

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

$$\rho_0 \in]-1, 0[\tag{3.30}$$

$$H_0 = 1 \tag{3.31}$$

$$P_0 = e^{-s\tau} \tag{3.32}$$

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_0 defined by Eq. 3.30 results from the demand that the reflex should be a negative feedback loop and that it must be stable.

In addition, it is assumed that also the transfer function P_1 of the predictive pathway represents unfiltered throughput given by:

$$P_1 := 1 \tag{3.33}$$

Finally it is assumed 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 like Eq. 3.14). Thereby, D also takes on the property of a typical transfer function.

Eq. 3.27 turns into:

$$\Delta\rho_j = \frac{\mu}{2\pi}\delta\rho_j \int_{-\infty}^{\infty} \underbrace{|DH_j|^2}_{A(i\omega)} \underbrace{\frac{-i\omega}{1-\rho_0 e^{i\omega\tau}}}_{-i\omega F(-i\omega)} d\omega \tag{3.34}$$

where $D = 1$ is set which represents a delta function as a disturbance.

Now Plancherel's theorem (Stewart, 1960) is applied to Eq. 3.34 to transfer the integral back into the time-domain and prove that it is negative. This leads to:

$$\Delta\rho_j = \mu\, \delta\rho_j \int_0^{\infty} a(t) f'(t) dt \tag{3.35}$$

The function $F(s)$ of Eq. 3.34 is given by the transformation pair:

$$F(s) = \frac{1}{1-\rho_0 e^{-s\tau}} \quad \leftrightarrow \tag{3.36}$$

$$f(t) = (-1)^n \delta(t-n\tau),\ n = 0,1,2,\ldots \tag{3.37}$$

where f represents an alternating delta 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 $f'(t)$, the temporal derivative of f.

The other term $A(s)$ of Eq. 3.34 is given by:

$$A(s) = |DH_j|^2 \tag{3.38}$$

$$a(t) = \Phi[d(t) * h_j(t)] \tag{3.39}$$

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

As a consequence of the above findings the integral in Eq. 3.35 has to be discussed which is specified by Eqs. 3.37 and 3.39. The integral should be negative to assure stability. From above it is known that D is short-lived with a duration shorter than τ, without which the loop-system would be instable to begin with. Thus, the discussion can be reduced to $t \approx 0$. It is known 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.

3.3.4.2 Real resonator-functions

Now real resonator functions for H_k and H_j are introduced (see Eqs. 2.2–2.5). Transfer functions of resonators are not orthogonal, but it will be shown by numerical integration that the system can still be treated as if orthogonal transfer-functions for H_k were used. In the case of non-orthogonal functions this results with (Eqs. 3.25,Eqs. 3.30–3.33 to:

$$\Delta\rho_j = \frac{\mu}{2\pi} \sum_{k=1}^{N} \delta\rho_k \int_{-\infty}^{\infty} \frac{-i\omega H_j^+ H_k^-}{1-\rho_0 e^{i\omega\tau}} d\omega \tag{3.40}$$

Fig. 3.3a shows the numerically obtained results for $\Delta\rho_j$ as defined in Eq. 3.40 in the case of a perturbation. Fig. 3.3b shows the equilibrium case with $\rho_0 = 0$.

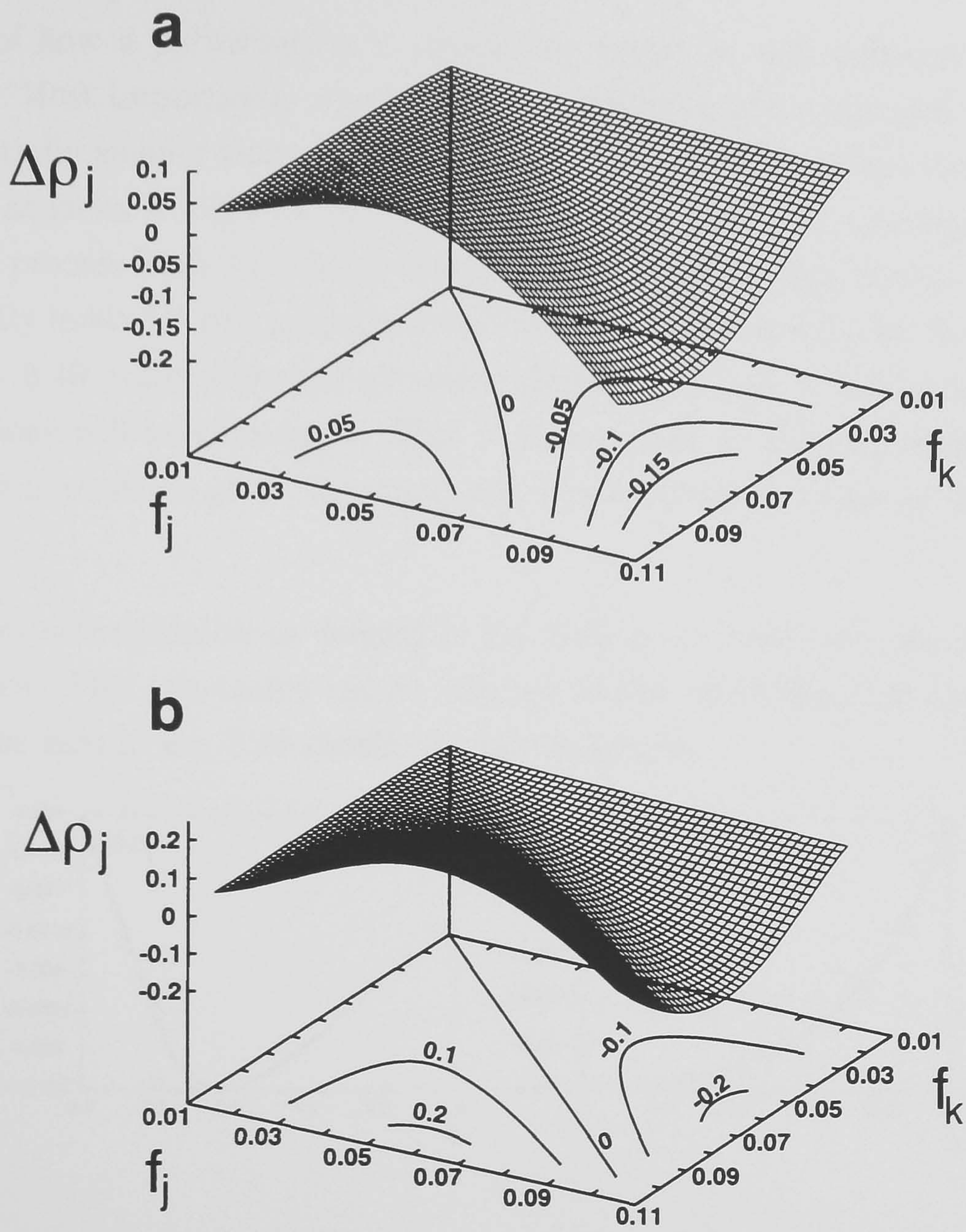

FIGURE 3.3: Numerical integration of Eq. 3.40. The disturbance was set to $D := 1$ and the delay τ was set to 1. The frequencies of the resonators (see Eq. 2.2–2.5) H_k and H_j were varied from 0.01 to 0.1 in steps of 0.001. The quality Q was set to 0.9. Part (a) shows the change of the weights ρ_j for $\rho_0 < 0$ and part (b) shows the change of the weights for $\rho_0 = 0$.

Note that the resonators are not orthogonal since for nearly all $j \neq k$ there are non-zero contributions. The system, however, still compensates for perturbations and, thus, converges, for the following reason. First, consider Fig. 3.3a, which represents the case of how the system values of the integral (Eq. 3.40) 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 those contribution has to be discussed and we have to 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. 3.3a) 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. 3.40, which describes all cross-interference terms, it can be argued 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, strict orthogonality as defined in Eq. 3.26 is not really necessary to assure convergence. This constraint can be relaxed to the constraint that the absolute value of the sum in Eq. 3.40 should remain below one.

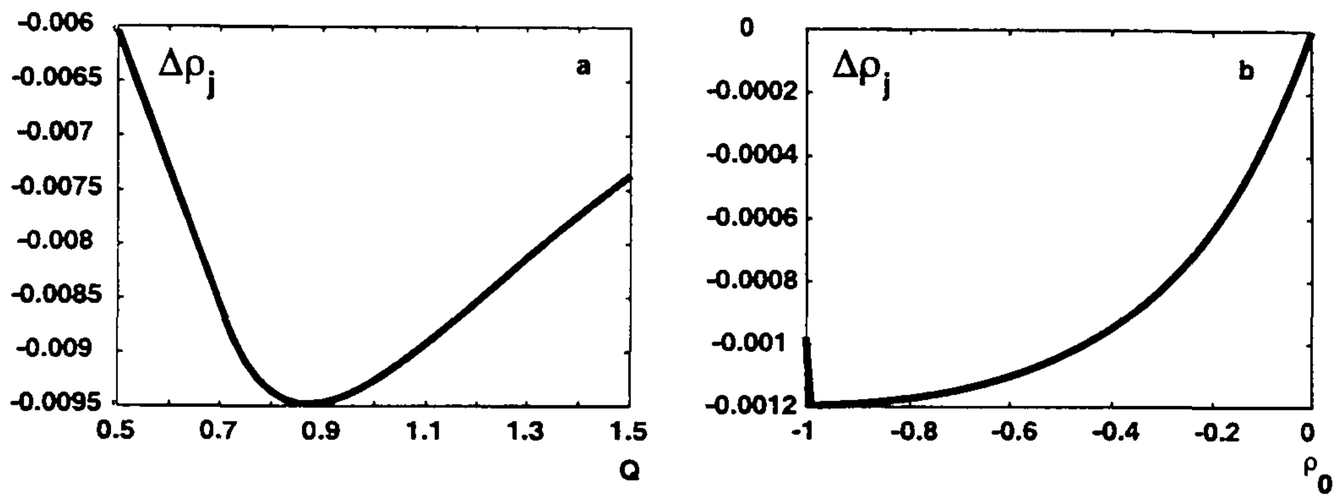

FIGURE 3.4: The best choice for ρ_0 and Q. Parameters: $f_k = f_j = 0.04$ for both plots and $q = 0.9$ for b).

To get optimal perturbation-compensation the diagonal elements in Fig. 3.3 should be kept as negative as possible. For that purpose the best values for ρ_0 and Q have to be found. Fig. 3.4 shows the result of integrating Eq. 3.40 for one diagonal element (see legend for parameters). The best value for Q is approx 0.85. The optimal solution for ρ_0 is at $\rho_0 \to -1$. This makes sense since the environment has a unity feedback and the case $\rho = -1$ is the limit where the compensation becomes unstable. A practical choice is below -1, for example 0.9 as used in Fig. 3.3. This result supports the limitations for ρ_0 (and in general for the feedback loop) which have been introduced at the beginning of this section with Eq. 3.30.

At this point the reader should be reminded of the introduction to the thesis: that from the beginning there must be a feedback which must "work" in the sense that it must be able to perform a specific task, namely to establish a desired state. In the context of this section it became clear also that the following learning behaviour

needs as a basis a working feedback loop to build up anticipatory structures. Thus, the general design principle is still first to build up an organism which has a *working* feedback loop and then give it the chance to build up anticipatory structures with the help of predictive learning[4].

3.4 Summary

This chapter has shown that ISO learning is able to turn a reactive system into a proactive system. The starting point in this chapter was therefore a reactive system. Such a reactive system has been introduced as a closed loop control system which is disturbed by an unpredictable disturbance. It reacts after a disturbance has caused a deviation from its desired state. To prevent such deviation from the desired state another sensor input is taken into account which is able to predict this deviation. It has been proven that ISO learning is able to use such a predictive input to generate an appropriate anticipatory action which eliminates the disturbance before it can cause a deviation from the desired state. In terms of engineering, ISO learning provides a forward-model of the reflex.

[4]From that point of view it would be interesting to leave the initial design of the feedback loops to evolutionary algorithms so that there is no need explicitly design them by hand.

Chapter 4

The Robot Experiment

4.1 Introduction

Up to this point ISO learning has been treated in a very general way without referring to any specific application. In this section ISO learning shall be tested in a specific application, namely in two robot experiments. These robot experiments use a more complex setup than in the theoretical derivations since the control of the robot demands more than one motor unit. Therefore the robot experiments will show not only the robustness of ISO learning but also suggest how to scale up to more complex situations.

The first experiment (section 4.2) involves a collision avoidance-task and the second experiment an *additional* attraction task (section 4.3). While the avoidance-task will show the robustness of ISO learning in a real world-task, the attraction- and avoidance-experiment will discuss observer-problems. It will be shown that the attraction experiment looks like a reward retrieval and that an observer is tempted to attribute internal reward-signals. However, it will be shown that there is no need for such a reward signal and that ISO learning solves the problem also by reflex-avoidance.

4.2 Avoidance reaction

The task in this robot experiment is collision avoidance. The avoidance experiment was first simulated on a computer in a simple environment containing a border and a few randomly placed obstacles. After this initial test-phase the program

was connected to a real robot via a standard I/O interface. The parameters were left the same as in the simulation using 10 ms time steps. The observed behaviour of the computer-simulation and that of the real robot were basically the same. In this section the data from the real-robot experiment is presented since it demands more from ISO learning in the sense of robustness than the simulation.

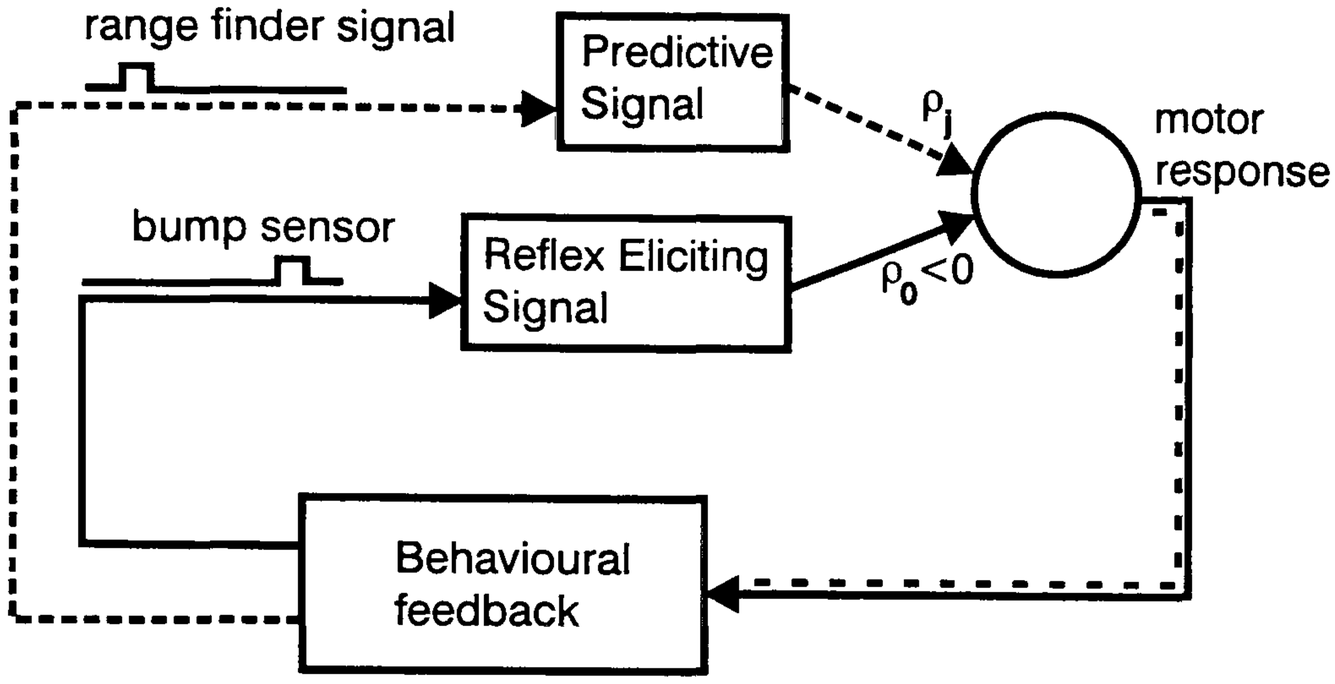

FIGURE 4.1: Simple sensor motor feedback with prediction which is made explicit by the example of collision avoidance. The solid lines depict a pre-wired reflex loop which exists before learning. This reflex loop performs a reflex reaction — in this case a retraction reaction (motor response) when the bump sensor (reflex eliciting signal) has been triggered. Learning has the task to learn that the earlier range finder signal (predictive signal, dashed pathway) can be used to generate an earlier motor reaction to prevent the bump (reflex).

The built in reflex behaviour is a retraction reaction after the robot has hit an obstacle (Fig. 4.1, solid pathway). This represents a typical feedback mechanism with the desired state that the signal at the bump sensor should remain zero. To prevent the robot leaving the desired state it can use other sensor modalities which can *predict* a looming collision. In our case this is achieved with range finders (Fig. 4.1, dashed pathway). The learning algorithm has the task of learning the existing temporal correlation between the range finder- and the bump sensor signals. After learning the robot can generate a motor reaction already in response to the range finder signals and thereby avoid the retraction reflex. Functionally, the reflex will be eliminated and the "predictive pathway" takes over after learning.

Up to this point the algorithm had been treated in a pure open-loop condition, where learning was entirely unsupervised. The robot experiments shown below create a situation where the behavioural reaction influences the sensor inputs, thereby creating a closed loop situation (Fig. 4.1). Unsupervised learning thereby turns into something which can be called "self-referenced" learning to distinguish it from "reinforcement" learning which requires an explicitly defined punishment

or reward signal, which is not present in closed loop ISO learning.

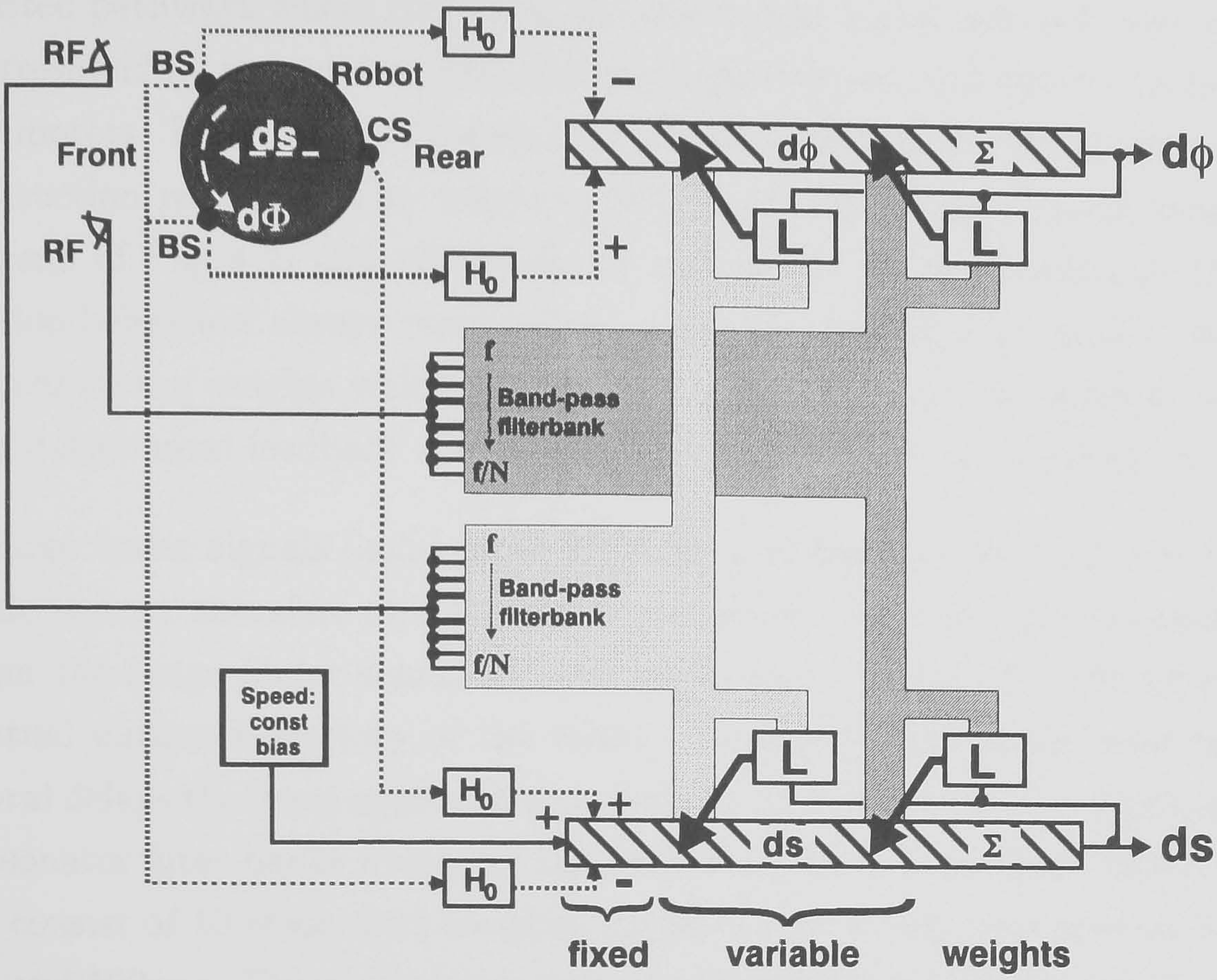

FIGURE 4.2: Robot-circuit: The robot consists of three bump sensors (BS), two range finders (RF) and two output neurons: one for the speed (ds) and one for the steering angle ($d\phi$). These output neurons represent simple summation circuits (indicated by $\sum$). The robot has a reflex behaviour which is established by the signals from the bump sensors (dotted lines) which are fed into 4 band pass filters H_0 with $f_0 = 1Hz$ and $Q_0 = 0.6$. The output of the band pass filters is summed at the neurons for speed (ds) and steering angle ($d\phi$). The corresponding weights are adjusted in such a way that the robot performs an appropriate retraction reaction if either of the bump sensors is triggered. The synaptic weights in this unconditioned reaction are kept constant at $\rho_0^{ds} = 0.15$ and $\rho_0^{d\phi} = -0.5$. The task of learning is to use the signals from the range finders (RF) to predict the trigger of the bump sensor (BS). The two signals from the left and the right range finder are fed into two filter-banks with $N = 10$ resonators with frequencies of $f_k = \frac{1Hz}{k}$; $k \geq 1$ and $Q = 1$ throughout. The 20 signals from the two filter banks converge on both the speed neuron and on the neuron responsible for the steering angle. Learning rate was $\mu = 0.00002$. L depicts the implementation of the learning rule (Eq. 2.6).

The robot's circuit diagram is shown in Fig. 4.2; a detailed description, which includes a list of the robot's control parameters is given in Appendix B. The robot has three bump sensors and two range finders. All signals are filtered by band pass filters and converge onto two neurons which generate two different motor outputs: one controls the robot's speed and the other the robot's steering angle. The speed of the robot is set to a fixed value and its steering to zero so that the undisturbed

robot drives straight forward. The built in retraction behaviour is generated by the dotted pathways where the bump sensors trigger highly damped sine waves in the corresponding resonators. This signal is sign-inverted and directly transmitted to the motors. Essentially, it consists of just one single half wave which, leads to the retraction reaction. The weights are initially set to appropriate values (see the legend of Fig. 4.2) and effectively do not change during learning so that the retraction behaviour always remains the same. The dotted bump sensor pathways with their strong weights which determine the motor output are together with the arising behavioural feedback equivalent to the reflex loop discussed in Fig. 4.1.

The range finder signals (solid lines) react at a distance of about 15 cm from an obstacle and are therefore able to predict a collision. However, the temporal delay between the range finder signal and the bump signal is variable and depends on the actual motion trajectory of the robot. To cope with a rather wide range of temporal delays the same approach as in section 2.5.2 has been used, implementing two resonator filter-banks which get their signals from the two range finders. Filter banks consist of 10 resonators covering approximally a temporal interval between $50\,ms$ and $500\,ms$. These resonator signals converge onto both the speed- and the steering neuron. Their weights are initially set to zero.

Depending on the initial conditions, different solutions were found by the robot to avoid obstacles. One solution, for example, is that the robot after learning simply stops in front of an obstacle or that it slightly oscillates back and forth. This type of behaviour may look trivial but is entirely compatible with the learning goal of avoiding obstacles. More commonly a different type of solution is observed where the robot continuously drives around and uses mainly its steering to generate avoidance movements. Other solutions do not seem to be possible and have not been observed. Furthermore, it must be mentioned that the robot always found one of these solutions after sufficiently long learning.

Figs. 4.3 shows episodes of the robot behaviour and its signals for one selected example trajectory. The signals shown in Figs. 4.3c,d correspond to a situation where the robot still collides with the walls. Corresponding collision points are marked in Fig. 4.3a by small letters c and d. As expected, learning leads to a change of the temporal relation between the range finder signal and the bump signal. This can be seen by the different lengths of T depicted in Fig. 4.3c,d and is due to the learned motor output which is increasingly dominated by the range finder signal. This supports the filter bank approach which has been used in the robot experiment. Finally, Figs. 4.3e depicts a situation where the robot

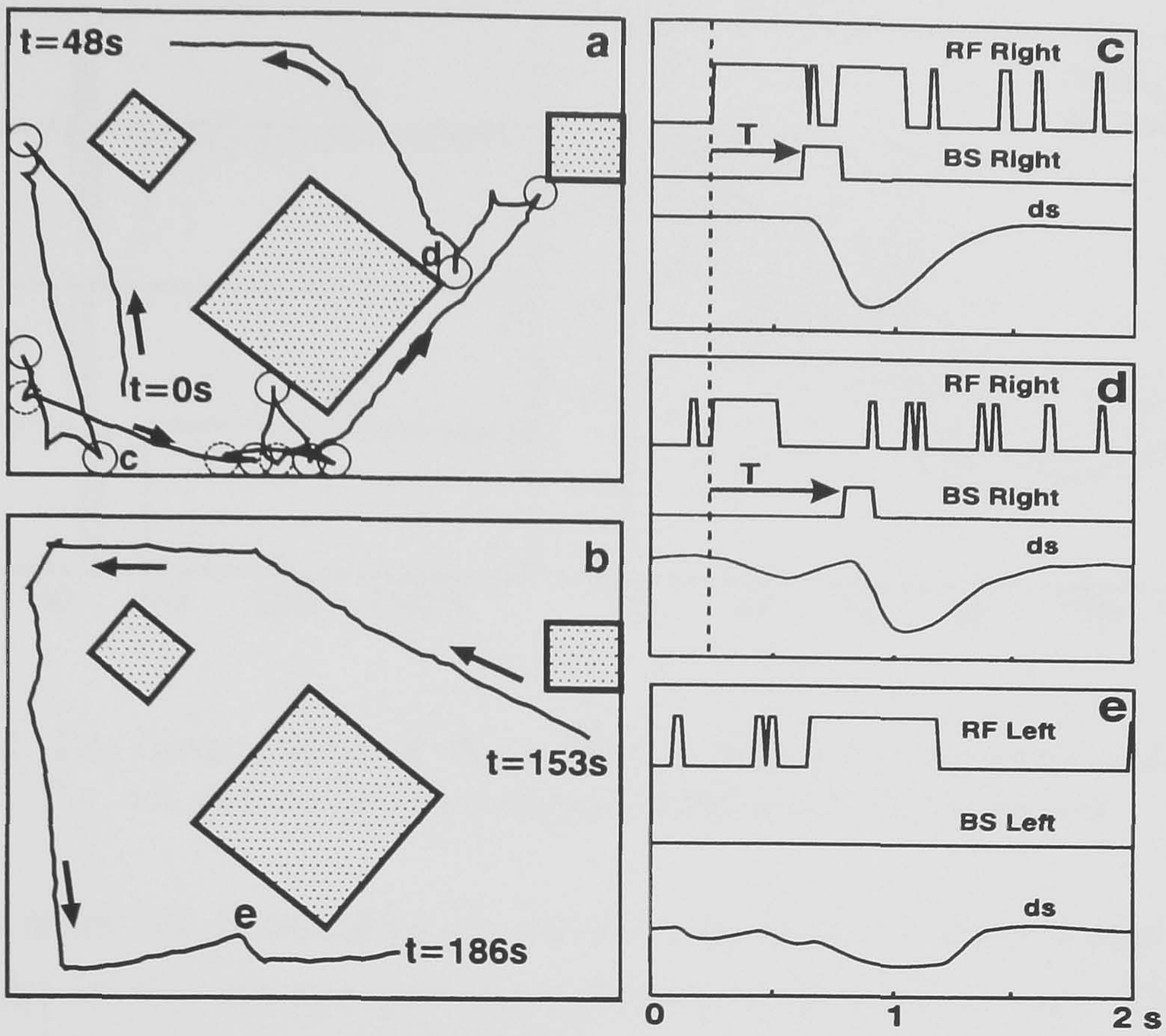

FIGURE 4.3: a) Manually reconstructed robot movement trace in an arena (240 $cm \times 200$ cm) with three obstacles (shaded) at the onset of learning. Motors were not entirely balanced leading to a curved start of the trajectory. Many collisions (circles, solid=forward-, dashed=backward collision) occur and trapping at obstacles happens. After a collision a fast reflex-like retraction&turning reaction is elicited. b) Robot movement trace after successful learning of the temporal correlation between signals at RF and BS. No more collisions occur, the trajectory is smooth. A complete movie of this trial can be viewed at `http://www.cn.stir.ac.uk/predictor/real` — movie 1 and on the CD which comes with this thesis (click on "avoidance learning"). c-e) Signals at RF (top), BS (middle), and motor control signal ds (bottom) for different learning stages. c) Signals occuring at the early collision marked 'c' in part a of this figure. A stereotyped motor reaction is elicited in response to the CS signal. d) Signals occurring at the late collision 'd'. Motor reactions occur in response to RF but are not sufficient to avoid the collision. When it occurs a strong motor reaction is again elicited. e) Signals occurring at the curve marked 'e' in (b). Smooth motor reactions occur in response to RF, CS remains silent because no collision occurs.

has learned to avoid the obstacles ($CS = 0$).

Note that the low pass component of the band pass filters smoothes the rater noisy range finder signals which substantially adds to the robustness of the algorithm. Furthermore, pure noise signals are not correlated to other sensor signals and do not contribute to learning.

The change of the weights in the robot example shall now be compared with

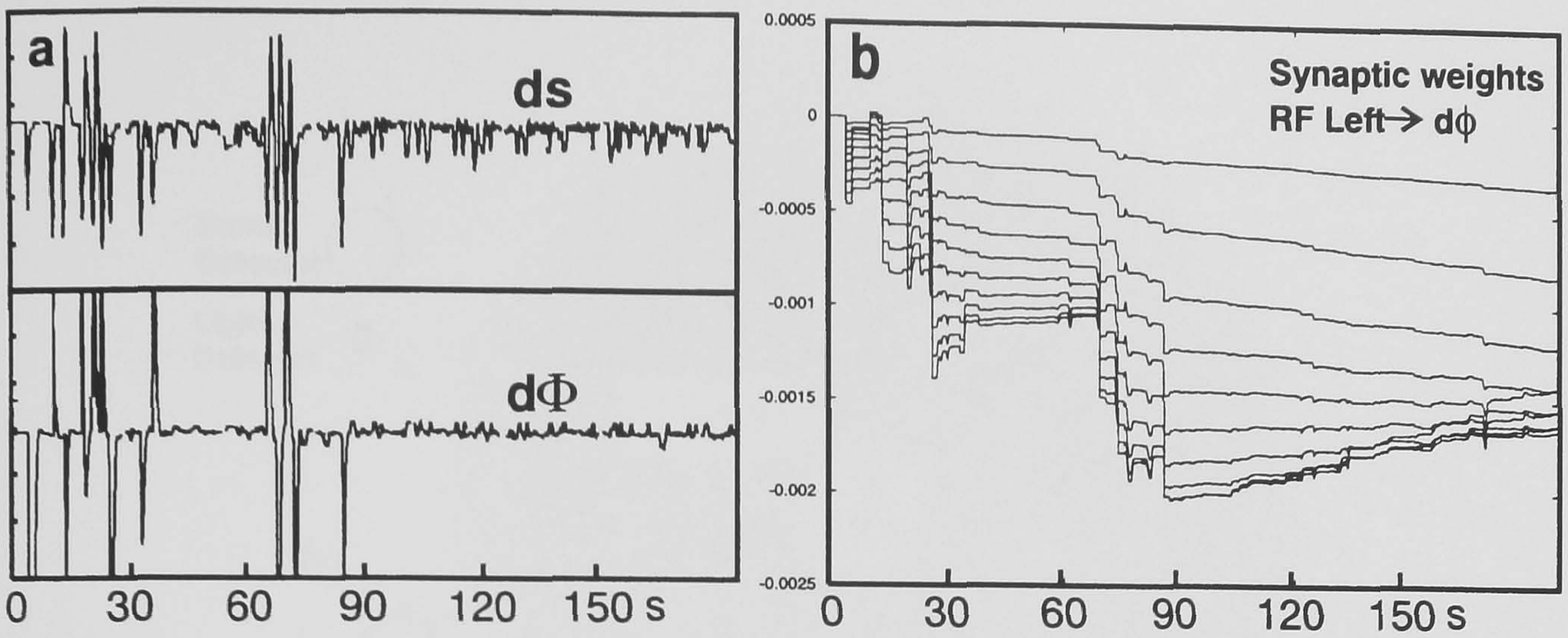

FIGURE 4.4: Development of the synaptic weights for the same trial as in Fig. 4.3 and the complete motor signal-traces for ds and $d\phi$.

the results from the simulations. It can be seen that the weights approximately stabilise also in the robot experiments presented here (Fig. 4.4). Their actual values depend on the solution found. The situation in the robot experiment, however, is more complicated than in the simulations shown earlier, because the $ds-$ and $d\phi-$neurons get signals from more than two sensors at the same time. Thus, very often triplets of temporal correlations exist, like during a slanted wall approach first a signal is obtained from the right, then one from the left range finder and finally that from the right bump sensor. After successful learning the bump sensor remains silent but the robot is still left with sequences of ranger-finder events. Thus, learning continues, though at a lower rate even after the last collision has happened (the bump has been avoided).

As a central observation, this shows that the system (here: the robot) continues to operate *without* a designated reference-signal (because x_0 is zero now). Learning continues between the remaining inputs (here: the range-finders).

This can, for example, be seen in Fig. 4.4 when looking at the development of the weight from the left range finder to $d\phi$ which continues to change after the last collision has occurred (at $t = 85\,s$). Ultimately, the earlier of the two range finder signals would dominate, but this will lead to a stable situation only for very simple (e.g., circular) trajectories where an unchanging relation between both range finder signals is forced upon the robot.

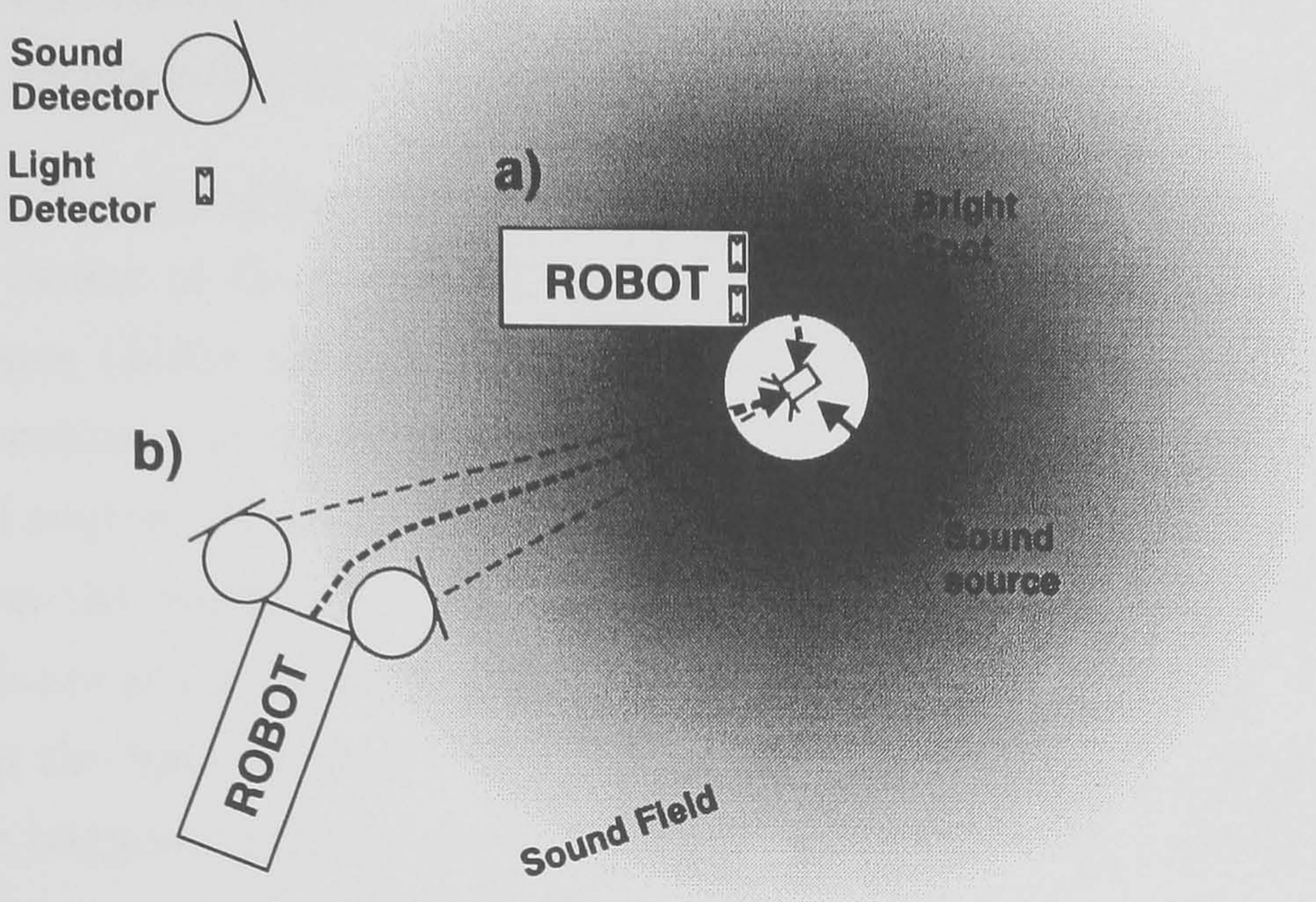

FIGURE 4.5: Illustration of the attraction reflex and the learned behaviour. a) When one of the LDRs enter the bright spot the robot drives to its centre which causes the spot to vanish. b) The two sound detectors SD enable the robot to locate the object from the distance.

4.3 Attraction- and avoidance reaction

The robot experiment of the last section showed only an avoidance reaction. In this section it will be shown as a computer-simulation that it is also possible to construct an attraction case with ISO learning. The computer-simulation presented here combines the avoidance- and attraction-reaction.

The design of the reflex reaction is the crucial point also in the attraction case. Therefore, the difference between an attraction reaction and an avoidance reaction is a different initial reflex reaction. While in the avoidance case the reflex is the avoidance of an object in the attraction case the reflex is simply the attraction of an object.

The reflex of the attraction case has the task to drive the robot towards the centre of a constantly illuminated area. At the moment the robot enters the centre the illumination vanishes and a new illuminated area appears somewhere else. This process could be interpreted as targeting and eating of food.

To establish this reflex the robot has been equipped with two light-dependent

resistors (LDRs). The signals of the LDRs are fed to the reflex input. The turning reaction is generated by using the difference of the signals between both LDRs (see Fig. 4.5) which causes a turn *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 from the centre of the illuminated area. The sound signals are detected by two microphones (MIC) attached to the robot. The difference of the microphone-signals provides a azimuthal information for the robot about the relative 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 turning-reflex. Learning stops as soon as the final turning-reflex is no longer needed. This is the case when both LDRs are triggered exactly simultaneously which means that the robot is heading straight for the centre of the illuminated area.

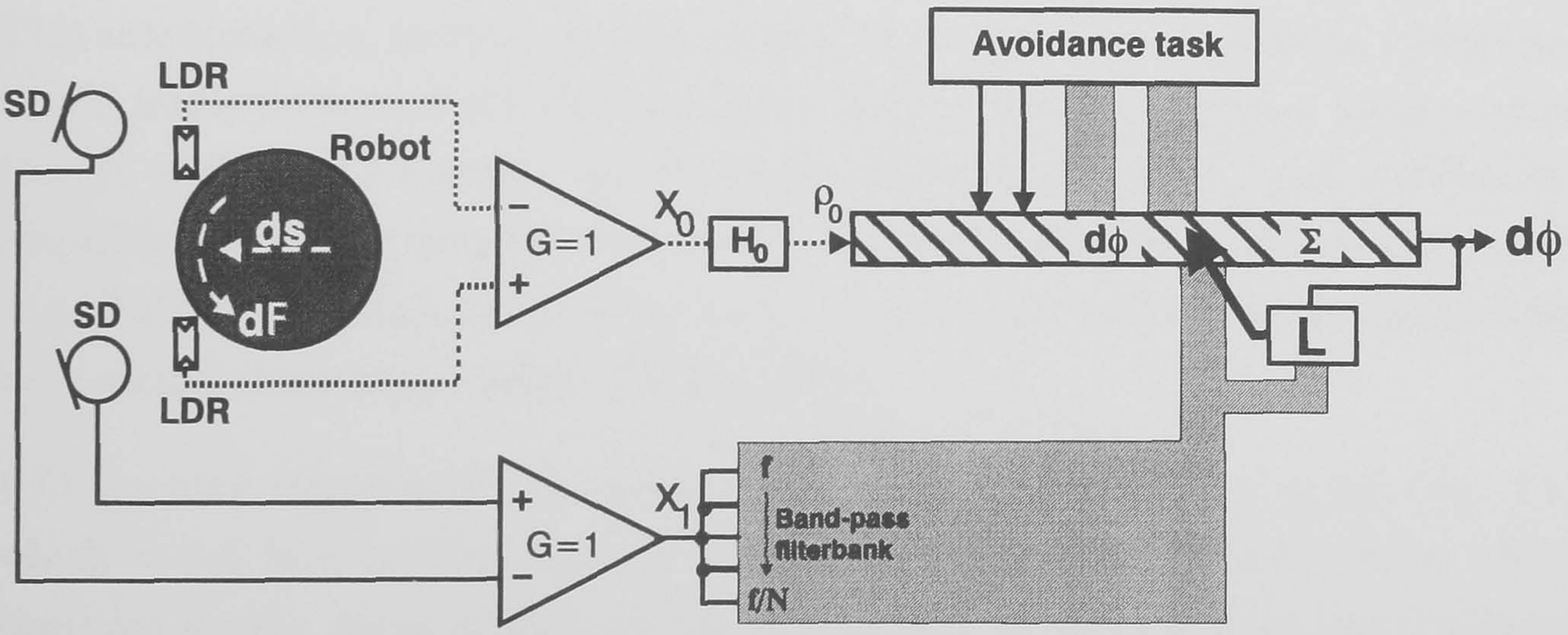

FIGURE 4.6: Additional circuitry for the computer-simulation of the attraction case. The avoidance circuitry is the same as shown in Fig. 4.2. The attraction task only involves $d\phi$. The light detectors (LDRs, signal-range: $[0 \ldots 1]$) establish together with the resonators $H_0\,(f = 0.01, Q = 0.51)$ the *attraction* reflex. The fixed weight for the reflex is set to $\rho_0 = 0.005$. The two sound detectors (SD) provide a signal which is inverse-proportional to the distance to a sound-source. The difference of the signals from sound detectors is fed into a filter-bank with $f_i = 0.1/i,\ i \in [1 \ldots 5]$ and $Q = 0.51$. The learning rate was set to $\mu = 0.0002$. All other parameters are identical to the avoidance task taking $10ms$ for one simulation-step.

Fig. 4.6 shows a more detailed view on the additional circuit needed for the attraction case which is based on the general circuit shown in Fig. 2.6b. The avoidance case is not shown but is identical to the circuit shown in Fig. 4.2.

The reflex reaction is triggered by the signals of the two LDRs which provide the

signal $x_0 = \mathrm{LDR}_l - \mathrm{LDR}_r$. The (here fixed) weight $\rho_0 > 0$ and the resonator $H_0\,(f_0 = 0.01, Q = 0.51)$ are arranged in such a way that the robot performs a stereotype turn *towards* the centre of the illuminated area.

The predictive signal x_1 is generated by using two MIC signals. The signal is simply assumed to give the euclidian distance ($r_{r/l \to m}$) of the left (l) or right (r) microphone from a sound source m. The difference of the signals from the left and the right microphone $r_{m\to r} - r_{m\to l}$ is a measure of the azimuth of the sound source m to the robot. Since there is usually more than one sound source in the playground the resulting signal is an average over all sound sources. Including a decay of the sound strength with the distance we get as the final difference signal for the M sound sources:

$$x_1 = \sum_{m=1}^{M} \frac{r_{m\to l} - r_{m\to r}}{\sqrt{r_{m\to l} r_{m\to r}}} \tag{4.1}$$

This mathematical model has been chosen to stay as close as possible to an electronic implementation for the real robot which includes a pulsed sound source ($8\,kHz$ with $62\,Hz$ bursts), two PLL-tone detectors (for 8kHz) and a difference amplifier which subtracts the averaged "lock-detect" outputs of the tone-detectors (e.g. XR2211). Female crickets use such chirps for the localisation of male crickets with similar frequency choices (Webb, 1995).

ISO learning receives at the predictive input x_1 this difference signal (Eq. 4.1) which is fed into a filter bank of 5 resonators with different frequencies which converge on the same $d\phi$-neuron which also gets the signals from the avoidance reaction. Learning is achieved as usual by Eq. 2.6.

Fig. 4.7 shows the trajectories before (a) and after (b) learning. Before learning the robot hits the illuminated areas by chance. At (1) entering the illuminated area causes a reflex-like reaction where the robot makes a sharp turn into the area. After such a turn, the area vanishes and a new one appears at another position. At (2), the robot shows the reflex-reaction after a bump. From time-step $24,500$ onward, no more illuminated areas are created so that their number decreases and after step $29,000$ the playground is empty. After learning it can be seen that the robot is directly targeting the illuminated areas and that it hits the areas now fairly centrally. This leads to the effect that no reflex reaction ($d\phi$) is caused when the robot enters the illuminated areas. Note that $d\phi$ itself can be non-zero as in (2). However, in all cases (1-4) $d\phi$ remains constant and therefore the derivative of $d\phi$ is zero. A zero derivative means that there is no learning and

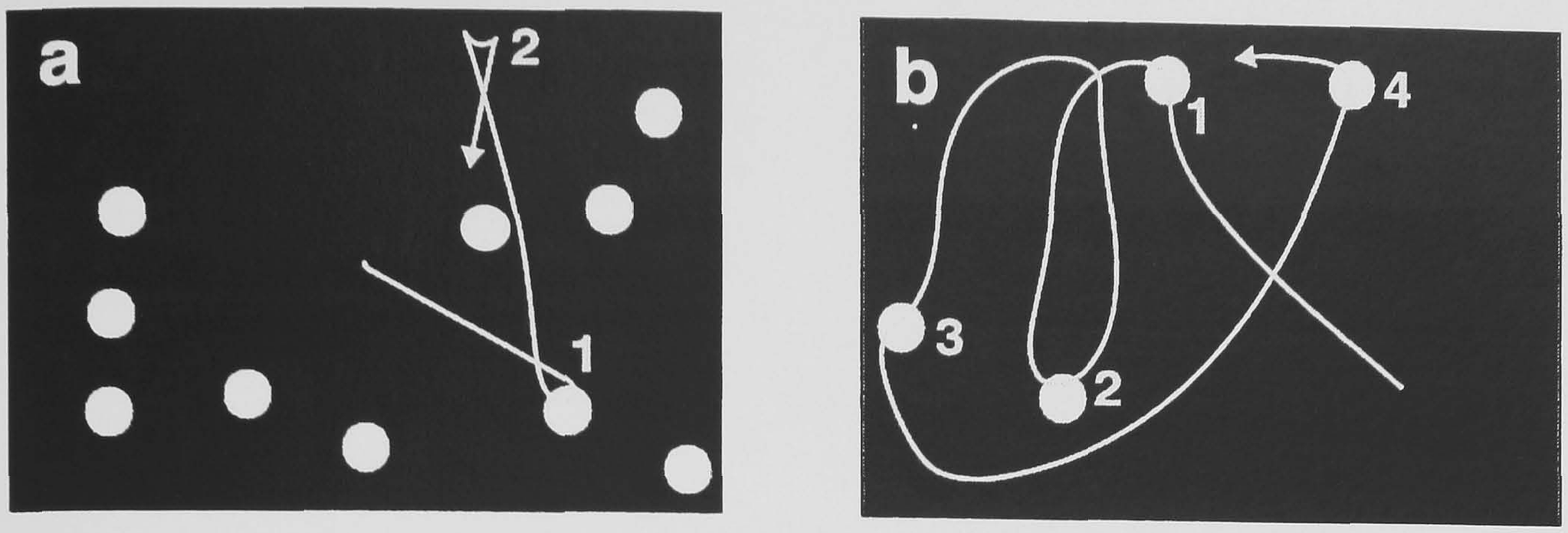

FIGURE 4.7: Trajectories of the attraction task before ($t = 0$) and after learning ($t = 21000 \ldots 24000$). a) before learning the robot randomly finds bright spots and bumps into the walls. Both the bright spots (1) and the walls (2) cause reflex reactions. b) after learning the reflex reactions have been replaced by an avoidance reaction on the one hand and by an attraction reaction on the other hand. Note that the robot's trajectory directly aims towards the centre of the bright spots. Therefore the robot enters the spots in a way that both LDRs are triggered at the *same* time (reflex is no longer triggered). The complete simulation can be seen at `http://www.cn.stir.ac.uk/predictor/animat/` and also on the CD (click on "Attraction and Avoidance Learning").

the weights ρ_j stabilise. This can be seen in Fig. 4.8 where the weights stabilise after approximately step $24,000$. At that point the playground is in the condition shown in Fig. 4.7b where the robot enters the illuminated areas centrally.

The weights turn out to be negative because of the setup of the LDRs and the MICS. For example, when the left LDR is triggered (which leads to $d\phi > 0$) the input to the filter bank is negative (the left microphone is closer to the sound source than the right one). Therefore the weights become negative.

Thus, it is also possible to construct an attraction-behaviour by ISO learning. Like in the avoidance case the initial reflex defines the attraction reaction. Learning the predictive attraction behaviour leads again to the situation that the initial reflex-reaction will be "*avoided*", in spite of the fact that this case deals with an *attraction*-behaviour (on the system-level of behaviour).

4.4 Summary

This chapter has presented two robot experiments which show that ISO learning is able to solve the classical obstacle avoidance task and that it is able to solve also an additional attraction task.

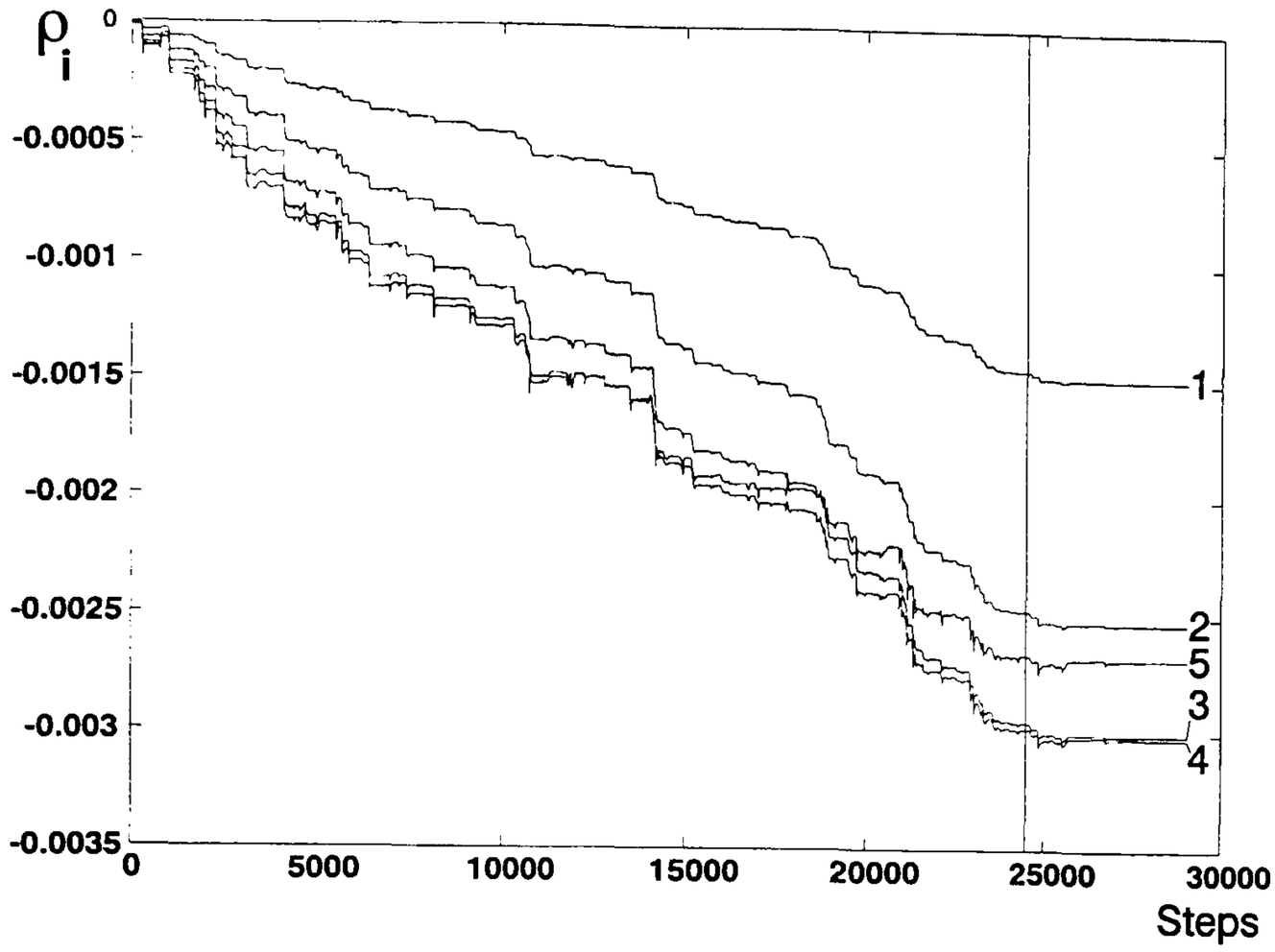

FIGURE 4.8: Development of the synaptic weights for the same trial as in Fig. 4.7. The dotted lines mark the area which is shown in Fig. 4.7b. The weight-index ρ_i corresponds to the index of the resonators: $f_k = f_1/k,\ k \in [1 \ldots 5]$

For the classical obstacle avoidance task the robot starts with a preprogrammed reflex: when the robot collides with an obstacle it retracts and then continues its journey. ISO learning was able to correlate the signals from the range finders with the trigger of the reflex reaction. The result was a turning reaction before the robot would collide with an obstacle. Therefore the collision avoidance experiment has shown that ISO learning is able to create an anticipatory action to avoid the trigger of the reflex reaction. This learning is very fast. Only a few collisions are needed to learn the avoidance reaction.

The attraction task defines an additional reflex reaction. In this experiment a playground was constructed with illuminated areas. The wiring of the reflex causes a turning reaction of the robot towards the centre at the moment when it enters such a illuminated area. If the robot is already heading towards the centre of the illuminated area no reflex is triggered. Such illuminated areas also emit sound signals which can be detected by the robot even from a distance. The sound-signals form the predictive signal which is used by ISO learning to generate a predictive reaction. This finally results with a behaviour whereby the robot heads to the centre of the illuminated area before it enters it. As a consequence, the reflex is no longer triggered. Thus, the proactive behaviour is here demonstrated by targeting the illuminated area from a distance and subsequently no reflex is needed when entering the area itself.

Finally, it should be noted that to the observer, the targeting towards the illuminated area from a distance it closely resembles a reward retrieval which seems to involve planning. However, it looks only like a reward retrieval. There is neither an internal signal which represents a reward signal, nor is there planning involved. ISO learning is always based on an avoidance learning, namely reflex avoidance. Therefore the observer must be cautious if he/she is interpreting behaviour and then attributing this behaviour to inner states.

Chapter 5

Discussing the Organism

5.1 Introduction

In the following sections ISO learning itself will be discussed in the open-loop condition. This means that the feedback from the environment is ignored and only the internal structure of the organism is discussed. Therefore this chapter refers to the results of chapter 2 which presented ISO learning without environmental feedback.

The discussion will start with similarities on the circuit level (section 5.2). An important part of ISO learning is the low-pass filtering of the input-signals. In technical applications it is common practise to low-pass filter signals and to utilise the properties of low pass filtered signals. This is especially the case in the field of Kalman filtering. Therefore the relation between Kalman filtering and ISO learning will be discussed.

Another similarity arises when the learning curves of ISO learning (see Fig. 2.2) are compared with learning curves of neurons. Physiological experiments have shown that the precise timing between input-spikes and output-activity determines the weight-change. The following section 5.3 will try to give the different parts of ISO learning a physiological meaning and will also point out the differences which arise when such an ANN[1]-rule like ISO learning is compared with physiology.

The last two parts of this chapter will discuss animal learning and will approach ISO learning from the level of behaviour (section 5.4). Animals change their behaviour while they are interacting with their environment. Also ISO learning

[1] Artificial Neural Network

changes the behaviour of an agent, especially from a reactive agent to a proactive agent. In animal learning there are two standard theories how learning is to be understood: classical conditioning and instrumental conditioning (also called reinforcement learning). Both learning schemes are related to ISO learning: both learn sequences of events by associating stimuli with each other. Classical conditioning will be discussed in section 5.4.1 and instrumental conditioning will be discussed in section 5.4.2.

Most of the psychological theories of animal learning are only on a descriptive level. However, there are mathematical models which try to model the behaviour of classical conditioning and instrumental conditioning. Therefore the remainder of the section about animal learning will discuss mathematical models of animal learning which are related to ISO learning (section 5.4.5). In particular the discussion will be guided by the distinction between algorithms which need an explicit reward signal and those which are able to learn without such a signal.

5.2 The predictability of low-pass filtered signals

ISO learning pre-filters all input signals (see Fig. 2.1) to render them predictable. The Kalman-filter which belongs to the class of adaptive filters also uses such a technique. This paragraph compares Kalman filtering and its underlying signal model with ISO learning.

One important goal of adaptive filtering is to separate unwanted noise from a signal (Bozic, 1979). For example, when an audio signal is being transmitted through a telephone line it will be disturbed by random noise. At the receiver there appears a mixture of the original signal and the noise. Adaptive filters have the task to strip off the noise and reconstruct the original speech signal. Another example comes from problems which are related to observation-processes. If one wants to track the trajectory of a plane one can achieve that with the help of radar (Bozic, 1979, p.136). Radar uses a rotating transmitter and receiver and one gets estimates of the plane's location at discrete time steps. Because of atmospheric disturbances or bad reflection one gets a noisy response which can only roughly estimate the plane's position. Thus, in both examples one receives a data-stream which is disturbed by noise and the goal is to reconstruct the original signal without the noise.

Having the two examples in mind they can be used to generalise to an abstract

model which describes an observation- or measurement-process. Such processes can be described by a mixture of the original signal (x_k) and the additive noise (v_k). Here discrete measurements are considered represented at k time-steps:

$$y_k = cx_k + v_k \tag{5.1}$$

where c is a constant and v_k is white noise with zero mean.

The task of adaptive filtering is to filter out the unwanted noisy components and preserve the original signal as much as possible. However, this is still too general since the signal-model of the un-disturbed signal x_k has not yet been defined.

The signal model of the undisturbed signal has to be defined as precisely as possible since it provides us with important a priori knowledge about the original signal. This knowledge makes it much easier to reconstruct the original signal from the disturbed signal. This becomes clear if the example of the plane is recalled. Consider the plane's trajectory: Since the plane usually cannot perform jerky manoeuvres and since the plane has a high inertia due to its mass one can conclude that the plane's trajectory is fairly *smooth*. More precisely, this means that the *future* development of the trajectory emerges out of its *past*. The trajectory can only be changed by a limited amount every time step and therefore it always incorporates the coordinates of the past. This also means that the trajectory is predictable for a certain amount of time. Such a signal model can be represented by the following time-discrete (k) recursive filter which gets white noise w_k at its input:

$$x_k = ax_{k-1} + w_{k-1} \tag{5.2}$$

where x_k is the output of the filter. White noise is a signal which does not depend on its past at all. Its autocorrelation function is zero (except at zero). The goal is to get a smooth and therefore *predictable* signal at the output of the filter. This is achieved by the recursive character of Eq. 5.2. The parameter $0 < a < 1$ determines how smooth (or how predictable) the output x_k shall be. With low values ($a \to 0$) the noise dominates and the output is only predictable for a few time steps. With high values for $a \to 1$ the signal becomes more and more dependent of its past and therefore more and more predictable[2].

The Kalman filter as one special case from the class of adaptive filters assumes the above signal model to separate noise from the original signal. The Kalman

[2]Compare to the so called "eligibility-trace" in TD-learning and in the Sutton and Barto-Model which is discussed later.

filter uses the a priori knowledge about the original signal (Eq. 5.2) namely that it is smooth and therefore its changes are slow and predictable. The Kalman filter gets at its input the signal x_k of the above recursive signal model (Eq. 5.2) disturbed by additive noise (see Eq. 5.1). The task for the Kalman-filter is to smooth out the disturbed signal y_k from Eq. 5.1 to eliminate the noise but without changing the shape of the original signal component x_k. Since the Kalman filter assumes that the signal is smooth and that it changes gradually it can be used to predict the course of the original signal x_k (usually one step ahead). This is a very important property used in many applications such as the previously mentioned radar tracking.

Up to this point the signal model of the Kalman filter has been interpreted in the time domain. However, the model can also be interpreted in the frequency domain. The above example of the telephone-transmission makes that clear. The original speech signal is band-limited. On the other hand noise is not band-limited. It spans a broad frequency-range. Thus, there are frequency ranges where there is only noise and there are frequency ranges where there is a superposition of the original speech signal and the noise. It is obvious that a filter which has its passband matched with the frequency range of the original voice signal eliminates the noise in an optimal way. The Kalman filter can be interpreted as such an optimal filter which filters the noise and preserves the original signal.

In the temporal domain the choice of the signal model for the original signal has been crucial for the success of the Kalman filter. The same applies for the frequency-domain. The signal model Eq. 5.2 now has to be interpreted in the frequency-domain. In the frequency domain Eq. 5.2 represents a first order low-pass. The choice of $|a| < 1$ adjusts thereby its cut-off frequency. The frequency distribution of the noise signal is flat. The filtering of the white noise by Eq. 5.2 leads to a distribution of the signal x_k which decays at higher frequencies.

From the discussion of the time domain, it becomes clear that the filter Eq. 5.2 renders the noise predictable. Thus, if one wants to turn unpredictable noise into a predictable signal the noise simply has to be low-pass filtered. This applies not only to noise. If the input signal is already smoothed out then the low pass filter makes it even *more* predictable.

In the frequency domain the a priori knowledge of the Kalman filter can be interpreted in the following way: the Kalman filter uses the property of the original signal, namely that it is band-limited. The Kalman filter operates as an adaptive low-pass filter which chooses automatically the optimal cut-off frequency which

is the cut off frequency of the filter of the signal model. In other words the cut-off frequency is chosen in a way that above the cut-off frequency there is only noise and below there is the superposition of noise and signal. This is the optimal solution which eliminates optimally the noise from the original signal.

The main difference between ISO learning and the Kalman-filter theory is that in the latter the operations are performed on the *same* signal (auto-correlation) whereas ISO learning calculates predictions between different signals (cross-correlation).

Like the Kalman-filter, theory ISO learning also makes use of the predictability of low-pass filtered signals. The input signals x_i are all filtered by the band-pass filters H_i. In contrast to the Kalman filter theory ISO learning does not assume that the input signals $(x_0, \ldots, x_N)$ are predictable. It must be stressed that ISO learning *renders* the input signals predictable and therefore makes use of the Kalman signal-*model* and not of the actual Kalman filter *theory*. The Kalman-filter theory is (implicitly) used by Der and Liebscher (2002) who state that the driving force of learning is to make the sensor-inputs themselves predictable. However, in the case of ISO learning there is no need to get smooth input-signals. The use of the derivative v' in ISO learning emerges from the fact that output signal v is smooth and that its derivative has a phase lead which can be used to employ predictive learning (Eq. 2.6). Thus, ISO learning makes the input-signal predictable whereas the Kalman filter-theory assumes that the original signal is predictable to reconstruct it.

Another difference between ISO learning and Kalman-theory is the actual filter setup. In the Kalman-theory the filter is an IIR-filter with variable coefficients which are adjusted during learning to the best cut-off frequency. In ISO learning the cut-off frequency is determined by a filter bank: Consider a noisy signal x_1 at the input of the filter bank (see Fig. 2.6). Each filter filters a different frequency range out of the input signal. Learning is achieved by Eq. 2.6 as a correlation of the filter-outputs u_k with the derivative of the output v. The question is now which input u_i leads to a weight change? The signals u_k which are uncorrelated with the output v (and therefore to any other input $u_j, j \neq k$) will average out at the end and the corresponding weight ρ_k will stay zero. Consequently, inputs which provide only un-correlated noise do not contribute to the output. Thus, ISO learning is able to filter out uncorrelated noise and to preserve the correlated signals at its inputs u_k. This is similar to the Kalman filter-theory in the sense that the noise is filtered out from the disturbed transmitted signal. However,

as stated above ISO learning uses correlations between different signals and the Kalman-filter theory uses *auto-correlations* of only one signal.

The low-pass characteristic of the transfer-functions is crucial to receive the phase lead in the derivative of v. However, there are other possible transfer functions for $H_0, \ldots, H_N$ which have a low-pass characteristic. Those possible transfer-functions shall be discussed here. The derivative v' is an integral part of the learning rule since it causes a phase lead in relation to the original function v. In the time domain this demands that the impulse response can be described in the ideal case by a pure sine-wave. This is the case with the resonator (see Eq. 2.2) which has a damped sine wave as impulse response. In the frequency domain there could arise different demands which are more determined by the actual application. For example, it might be useful to filter out DC-components from the input signals. The simplest solution would be to introduce a zero-crossing in Eq. 2.2 so that the frequency response at $\omega = 0$ would be forced to zero. However, this is not possible since the transfer function

$$H(s) = \frac{s}{(s+p)(s+p^*)} \tag{5.3}$$

contains a phase lead

$$h(t) = \frac{1}{b} e^{at} \left[b \cos(bt) - a \sin(bt) \right] \tag{5.4}$$

in form of a cosine. Thus, if one wants to filter out DC-components one has to do this by a high-pass in front of the resonator. This high-pass cannot be an ideal high-pass $H_{\text{high}} = s$ since it would again lead to Eq. 5.3. A solution would be to use a real high-pass with a non-zero cut-off frequency. Such a high-pass must be designed in a way that the demand of a phase-lead is still not violated.

5.3 Mapping ISO learning to neurophysiology

The remainder of this paragraph will explore how close the learning rule (Eq. 2.6) is related to neurophysiology. The most striking similarity between ISO learning and neuronal plasticity can be established in the field of spike timing dependent plasticity (STDP). The common feature of STDP is that the *timing* of the post- and presynaptic activity determines the actual change of the synaptic weight. If the presynaptic activity precedes the postsynaptic activity the corresponding weight increases and if the timing is reversed the weight decreases. This type of

plasticity has been explored in the tectum, the hippocampus and also in the cortex of several species (Markram et al., 1997; Zhang et al., 1998; Bi and Poo, 1998; Xie and Seung, 2000). Abbott and Nelson (2000) and Bi and Poo (2001) summarised the different aspects of STDP in the different brain regions. These observations have been formalised by Gerstner et al. (1997), Kistler and van Hemmen (2000), Song et al. (2000), and Song and Abbott (2001) in a spiking neuron model. A review about the theory of synaptic plasticity in spiking neurons can be found in van Hemmen (2001). Fig. 5.1 shows one example of a recorded learning curve of

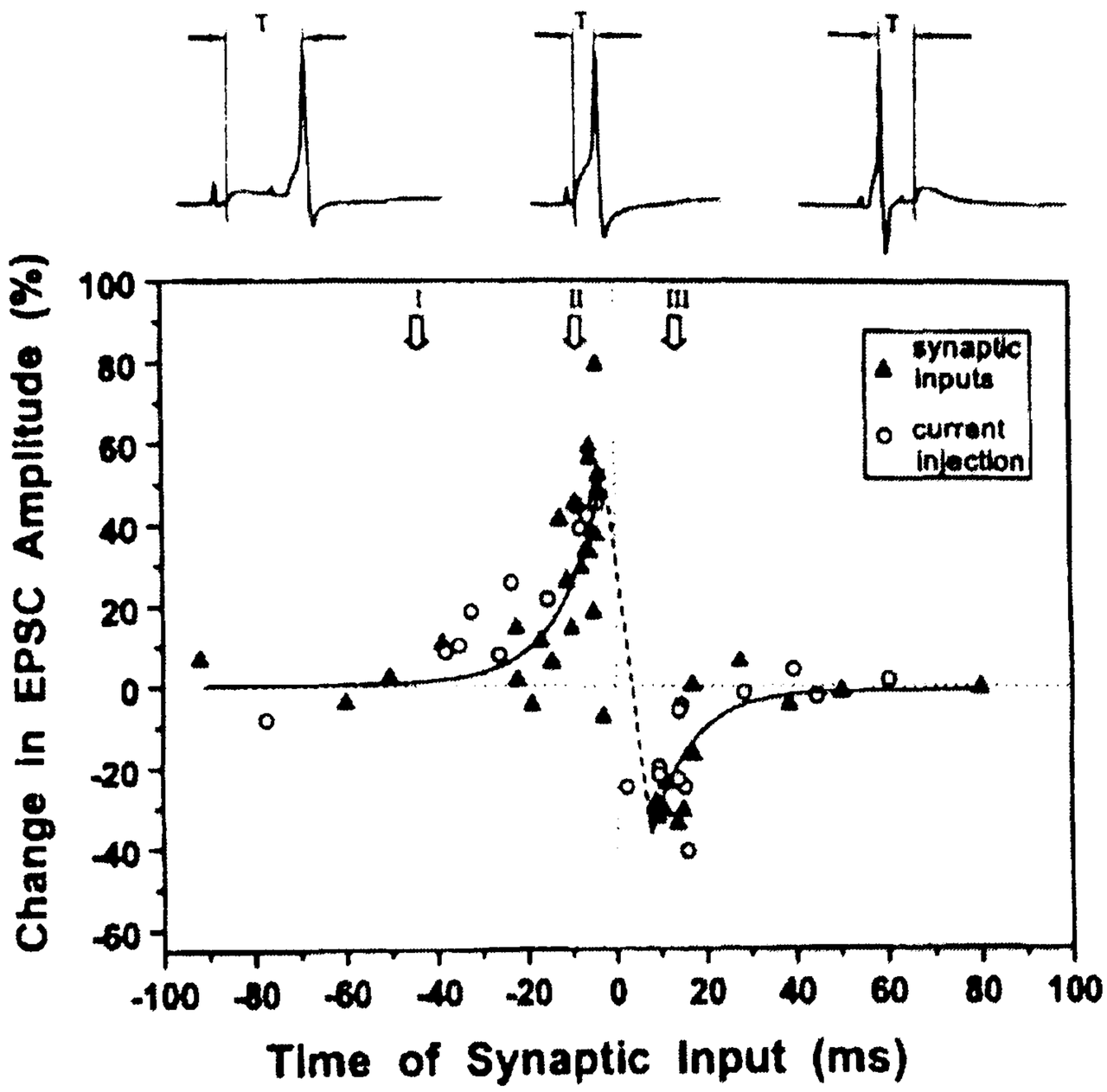

FIGURE 5.1: The learning curve of tectal neurons from Xenopus. The graph depicts the resulting change of excitatory postsynaptic potentials dependent of the timing of pre- and postsynaptic activity (Zhang et al., 1998).

spike timing dependent plasticity. The learning curve is similar to the learning curve shown in Fig. 2.2. Especially if identical resonators ($H = H_0 = H_1$) with identical qualities ($Q = Q_0 = Q_1$) are used the resulting basic shape is similar to the learning curves taken from neuro-physiological data.

There are two ways to establish a link between the neuro-physiological data and ISO learning. The difference arises from the *interpretation* of the signals in ISO learning. The signals can be either interpreted as *membrane-potentials* or, on the other hand, they can be interpreted as *firing-rates*. Firing rates can be seen

as a linear first oder approximation of a spiking neuron model (Kistler and van Hemmen, 2000; Song and Abbott, 2001). The rate-code shall be discussed first.

The interpretation of the signals in ISO learning as rate codes made it possible to develop analytical solutions in the closed loop case. Although the timing of single spikes is lost in the approximation of a rate-code the link to STDP is still existing. Thus, there is still the opportunity to develop a learning rule which is dependent on the pre- and post-synaptic timing although only a rate code is employed. This has been shown by Xie and Seung (2000), and Roberts (1999). Both establish a link between the STDP learning curve and the learning rule Eq. 2.6. Xie and Seung (2000) do not assume a specific STDP learning curve. To get firing rates they averaged the post- and presynaptic spike trains over time and calculated the cross-correlation function of the post- and presynaptic firing rates. Then, they developed the cross-correlation functions into a Taylor series. This finally resulted in the following learning rule:

$$\dot{\rho}_{jk} \propto \nu_k[\beta_0\nu_j + \beta_1\dot{\nu}_j + \ldots] \tag{5.5}$$

where:

$$\beta_0 = \int_{-\tau}^{\tau} f(u)du \tag{5.6}$$

$$\beta_1 = \int_{-\tau}^{\tau} uf(u)du \tag{5.7}$$

ν_k is the presynaptic firing rate, ν_j is the postsynaptic firing rate, ρ_{jk} is the corresponding weight, $f(u)$ is the learning curve and τ determines the limits for the integration. The learning curve $f(u)$ determines the weight change between a pair of pre- and postsynaptic spikes which have a temporal difference of u. In the case of classical Hebbian learning $f(u)$ is symmetric at zero ($f(-u) = f(u)$) and in the case of STDP $f(u)$ is antisymmetric ($f(-u) = -f(u)$). It is interesting that in the case of a symmetric learning curve Eq. 5.5 becomes the classical Hebbian learning rule

$$\dot{\rho}_{jk} \propto \beta_0\nu_k\nu_j \tag{5.8}$$

and in the case of an antisymmetric learning curve, Eq. 5.5 becomes the so called differential Hebbian learning rule:

$$\dot{\rho}_{jk} \propto \beta_1\nu_k\dot{\nu}_j \tag{5.9}$$

which is equivalent to the learning rule presented here (see Eq. 2.6). Thus, a

completely antisymmetric learning curve between spiking neurons also leads to temporal Hebb (or differential Hebb) in a rate-coded model. Roberts (1999) also applied a Taylor series, however, more explicitly in using the derivative of a gauss-distribution for the learning curve $f(u)$. This results in a learning rule which contains the derivative of the postsynaptic spike-probability as a multiplicative factor. The pre-synaptic activity can enter the learning rule after any transformation (which includes the unmodified pre-synaptic rate).

Therefore rate-coded models can account for spike-timing dependent plasticity if the corresponding learning rules contain the derivative of the output-rate of the cell. Roberts (1999) argues that the derivative of the postsynaptic rate should be taken into account as well as the derivative of the presynaptic rate. In ISO learning this is not possible. To stabilise the weights the correlation of the signal with its derivative is needed (Eq. 2.6) or in other words: the correlation between a sine and a cosine.

The main advantage in contrast to the above spiking models is that a rate-coded model can be treated analytically with the help of signal/control-theory and it is easy to integrate the environment in the model. In spiking neuron models, the underlying mathematical description is usually based on *statistics*. This makes it extremely difficult to deal with system-theoretical models which involve more than one processing step like the closed loop model presented here. Another argument to use a rate code comes from the aspect that ISO learning directly transfer sensor signals into motor outputs. Sensor and motor surfaces usually rely only on rate codes (Shepherd, 1990, pp.32–66). Since sensor signals are directly transferred into motor signals one can justify a model-neuron which operates with rate codes.

The interesting difference to spiking neuron models is the origin of the learning curve. In spiking neuron models, the learning curve is generally a fit to neurophysiological data. The actual function is usually a difference of exponentials to allow for an easier statistical treatment (Kistler and van Hemmen, 2000). In ISO learning the learning curve *results* from the impulse-responses of the resonators (H_k). In other words: the input dynamics determine the shape of the learning curve.

Now the low-pass characteristic of the resonator H_k has to be discussed in the context of neurophysiology. Low pass characteristics are very common in neurophysiology. They reflect the fact that any system needs a certain time to react. Therefore low-pass filtering can be found as a basic property of nerve cells (leaky integrator), in many receptor responses and in the change of chemical potentials

(Shepherd, 1990, pp.32–66). Thus, such low-pass characteristics are also one of the basic properties of any neuronal cell model (Koch and Segev, 1989).

Now, the signals in ISO learning shall be directly identified by signals in a spiking neuron. In principle, there is no obstacle in transferring ISO learning to a spiking neuron-model. The test-signals at the inputs $x_0, \ldots$ of ISO learning (see Fig. 2.1) have been delta-pulses and could easily identified as pre-synaptic spikes (Rieke et al., 1997, pp.281–283). The problem arises from the actual mapping of cell-properties to ISO learning. The low-pass characteristics seem to be no obstacle. Especially STDP is strongly linked to the dynamics of the NMDA channel (Ekström et al., 2001) which exhibits the right timing properties. Thus, it should be relatively straightforward to redesign ISO learning into a biophysically more realistic one, which directly relies on such internal neuronal variables and which uses spike trains as inputs. However, the identification of the derivative in neurophysiology is much more difficult and still poses some problems. Another form of sequence learning (TD-learning) uses also the derivative of the output-signal. Mapping TD-learning onto neurophysiology has recently been attempted by Rao and Sejnowski (2001) using the TD-learning algorithm but the relation between TD-learning and STDP is less direct and, accordingly, the transition between those two models is bit more intricate (Dayan, 2002). Especially the mapping of the derivative to neurophysiology was not successful.

At this point the learning between different inputs of the neuron has to be considered and it has to be discussed between which inputs learning takes place[3]. Synaptic potentiation in biology usually happens under homo-synaptic learning (Bi and Poo, 2001). This means that a synapse is potentiated when it receives *both*, a pre-synaptic spike and a postsynaptic spike. Thereby the pre-synaptic activity has to precede the postsynaptic activity (by approx $5\,ms$) (Nishlyama et al., 2000). If the timing is reversed depression is induced. Hetero-synaptic learning changes a synapse which only gets a postsynaptic spike but not pre-synaptic one. Heterosynaptic LTP usually does not happen in biology (Nishlyama et al., 2000) but in rare cases it has been observed (Bi and Poo, 2001). No heterosynaptic learning happens in ISO learning: Eq. 2.6 does not change the weight ρ_j when the input signal u_j is zero.

In section 2.5.1.4 the change of the weight ρ_0 has been discussed. In some cases it was desirable that ρ_0 stays constant, especially if ρ_0 is strong and represents

[3]The ISO learning rule defines its goal at the input and not at the output. There is at least a weak link in the work by Anastasio (2001) who explains the VOR-reflex with a cerebellar model which involves no error signal but the minimisation of the overall *input* at the purkinje cells.

an important reflex which should not vanish during learning. To keep ρ_0 constant some additional measures have been suggested. Here it can be shown that in biology this problem often does not exist. In biology there is the possibility given that strong weights stay stable even if the STDP suggests a small decrease of the synaptic strength (see Fig. 2.5 where ρ_0 slightly decreases). The stabilisation of the strong weight can be achieved by homo-synaptic self-potentiation of the same pathway. Potentiation in general is strong in weak synapses and weak in strong synapses (Guo-Quing and Poo, 1998). However, the synapse must at least be strong enough to cause a postsynaptic spike when it is triggered by a pre-synaptic spike because potentiation only happens in conjunction with a postsynaptic spike. Thus, once a synapse is strong enough to cause a postsynaptic spike it will maintain its strength by itself. ISO learning has the problem that if one allows all weights to change the weight of the reflex ρ_0 will slowly decline (Eqs. 2.21,2.22). This may be an unwanted effect especially if the reflex is essential for survival and the weight ρ_0 has to be kept constant in ISO learning. However, in biology this problem does not seem to exist since the decline of the weight ρ_0 will be compensated by homo-synaptic potentiation.

There is a variety of cases when an organism has to navigate successfully through a spatial area. The goal might be to find food (Blum and Abbott, 1996) or to avoid objects. To navigate successfully the organism must have spatial information about its environment. Neuro-physiological data supports this assumption in that some animals have specialised cells which fire when the organism is at a specific *place*. Therefore these cells are called "place cells" (O'Keefe, 1976; Ekström et al., 2001). When a rat is exploring a maze such a place cell fires — after learning — at a specific place in the maze. Since the rat encounters different places in the maze different place cells fire at different times. Therefore the place cells fire in a temporal sequence. Such a temporal sequence is ideally suited for temporal sequence-learning algorithms like ISO learning or TD-learning. There have been successful attempts to use the output of place cells for TD-learning (Arelo and Gerstner, 2000) to learn to find a target.

ISO learning can also use the input of place cells to learn a sequence of places. The reference is again the reflex behaviour which has nothing to do with the place cells. It can be implemented in the same manner like in the attraction- or avoidance-examples (as a retraction-behaviour or as a movement towards a light-source, ...). The place cells can be used as the predictive inputs of the learning circuit $x_1, \ldots, x_N$. Learning starts at the place cell which fires directly before the reflex reaction is triggered. This place cell becomes the predictor for the final reflex

and can trigger the earlier anticipatory behaviour. Once this behaviour has been learned a second place cell can predict the first one and so on. Thus, the temporal sequence which is formed by the place cells is learned and used to generate an anticipatory response caused by the earliest place cell.

Place cells provide a convenient form of pre-processing of the raw visual input for ISO learning (and for other temporal sequence learning algorithms). They generate from the intensity-levels of the visual field a temporal sequence of events. In general it can be stated that often a certain form of pre-processing of the inputs is desirable so that a sequence of events is generated when the organism is moving in its environment. Place cells can be an appropriate form of input to our algorithm.

5.4 Animal learning

In the next sections learning paradigms shall be explored which exist in psychology and animal behaviour (Mackintosh, 1974), and they shall be compared to ISO learning. It is important to keep in mind that ISO learning is designed for a closed loop learning-paradigm and that in psychology this differentiation between open and closed loop is often not used. Therefore direct comparisons to ISO learning in the sense of benchmarks are not possible. Additionally one should keep in mind that in constructivism attributions towards internal states are not permitted because they do not reveal themselves to the observer. Therefore the construct of the "reward" can not be used since it associates behaviour with internal states. The first topic is classical conditioning.

5.4.1 Classical conditioning

One of the oldest paradigms of animal learning is classical conditioning (Pavlov, 1927). The standard example of classical conditioning is Pawlow's dog which salivates when it gets food. This is the unconditioned reaction (UR) namely salivation to the unconditioned stimulus (US) "food". If the sound of a bell precedes the food the dog starts to salivate when it hears the sound of the bell. This is the conditioned reaction (CR) to the conditioned stimulus (CS). Note that the unconditioned reaction and the conditioned reaction are the same.

Any feedback from motor output to sensor input is usually ignored or explicitly

interrupted in classical conditioning (Domjan, 1998)[4]. The remainder of this paragraph will attempt to give some arguments against the open-loop assumption of classical conditioning. It seems not to be realistic in the light of a feedback/feed-forward system theory.

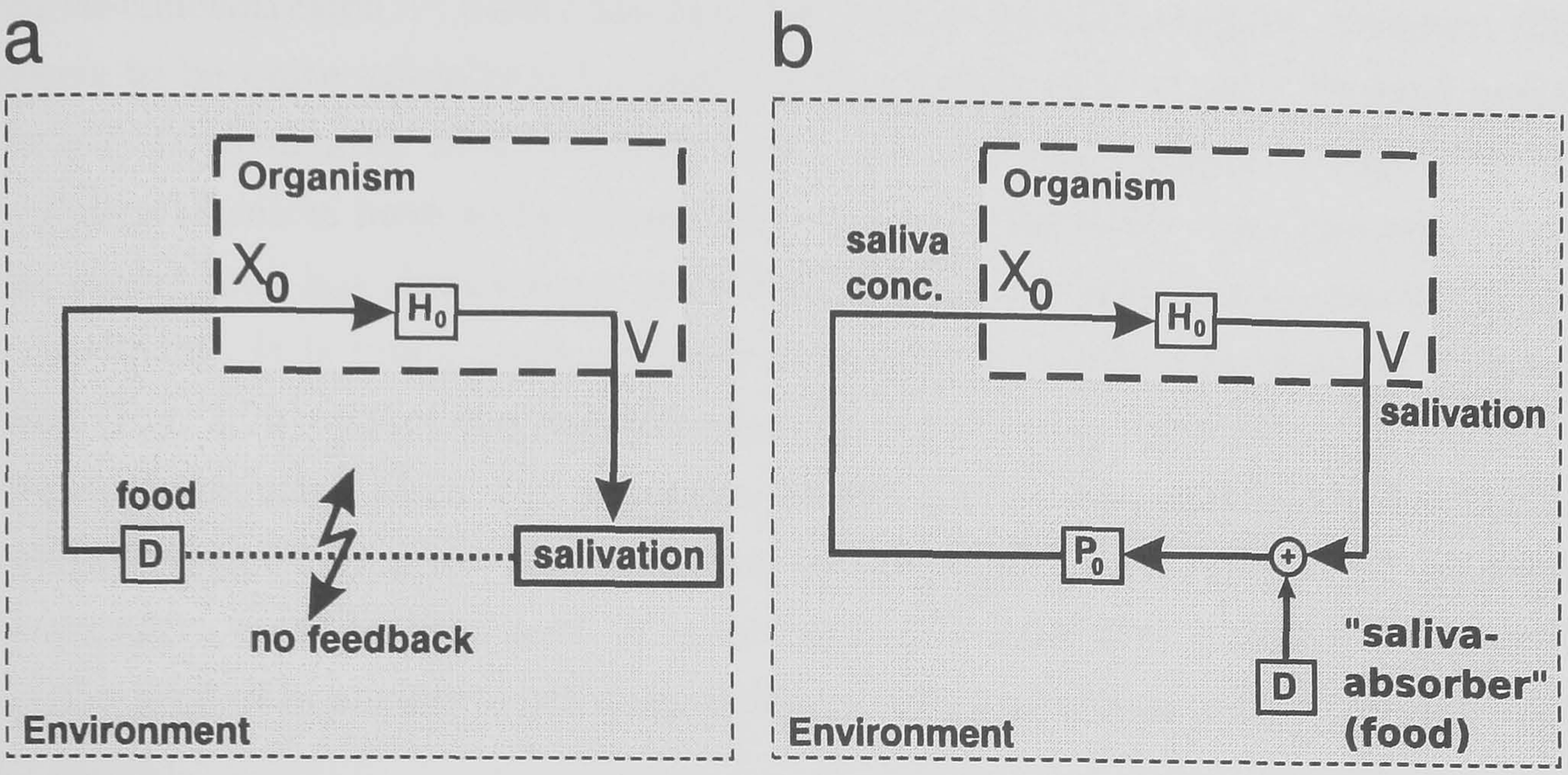

FIGURE 5.2: Pavlov's experiment seen as a open-loop experiment (a) and as a closed loop experiment (b). a) Open loop case: the food triggers the salivation reaction which does *not feed* back to the sensor input. b) Closed loop case (self-referential system with saliva as elements): saliva is absorbed (by food) and causes a lack of saliva in the throat. The lack of saliva is detected by a saliva sensor in the throat and triggers production of saliva. The saliva production *feeds* back to the saliva detector.

The focus shall be on the dog's reaction (UR and CR), namely the salivation (see Fig. 5.2). The experiment with the dog shall be interpreted both in a closed loop paradigm and in an open loop paradigm. In classical conditioning the act of salivation does not change the sensor input(s) — in this case the smell of the food and/or the sound of the bell (Fig. 5.2a). This means that the dog's action (salivation) does not feed back to the dog's sensor inputs (Domjan, 1998, inner cover). For the dog this means that it does not know if the salivation has any effect in his mouth. This finding can be analysed in the light of the feedback/feed-forward paradigm (compare Fig. 3.2 and Fig. 5.2). In the context of this paradigm it seems to be that the dog is using a *forward model* for his salivation since there is no motor-sensor feedback. This corresponds to the outer pathway in Fig. 3.2 via D, P_1, X_1 and V with $P_{01} = 0$. As pointed out in section 3 the exclusive use of the outer pathway has been reached if H_V has become the inverse of the environmental

[4]For example, Domjan is one author who explicitly elaborates about the open-loop character of classical conditioning.

transfer function which means that the organism has complete *knowledge* about the environment concerning the specific task (salivation, obstacle avoidance, ...). If the general findings of section 3 are related to Pavlov's experiment the consequence is that the dog already knows exactly how much to salivate to get the right saliva-concentration — before the food has even entered the mouth. However, this seems to be quite unrealistic because the dog would have to use a forward model already implemented as a pre-wired reflex. Food is not uniform so that different amounts of saliva have to be produced to make it digestible. One can argue that the closed loop has simply been ignored (Lieberman, 1993) by the observer of the experiment. It is much more probable that the salivation is involved in a closed loop (Fig. 5.2b) where the concentration of the saliva is measured by a sensor in the dog's mouth. Thus, the salivation might not be triggered by the food, it is rather triggered by a low concentration of the saliva in the mouth and therefore more saliva must be produced. It is interesting to note that in Fig. 5.2b develops on the level of behaviour a self-referential model on the basis of saliva (saliva leads to saliva).

Changing the interpretation of the experiment from open-loop to closed loop leads also to the change of the unconditioned stimulus (US). In the original example by Pavlov (open-loop) it is the food which triggers the salivation (see Fig. 5.2a). In the closed loop interpretation (see Fig. 5.2b) it is the concentration of saliva in the mouth. The unconditioned stimulus (US) can be a saliva-sensor which measures its concentration and triggers salivation if its getting dry (Lieberman, 1993). Learning any predictive cue, like the bell, can start the salivation earlier and can prevent the unwanted situation when the food is being eaten and the mouth is not yet wet.

As stated in the introduction an organism can only rely on feedback mechanisms (including internal feedback like, for example, memory). However, the observer looks from the outside at the organism and does not see the feedback loops of the organism since for the observer it is tempting to treat the organism as an input-output system to integrate the organism in his/her own feedback loops. However, even if the external observer is aware of his/her observer status it is difficult for the external observer to identify the feedback from the motor output to the sensor input. Only the behaviour can be observed but which sensor inputs are used cannot be observed.

The observer-perspective leads to another problem: the interpretation of the food as a reward. Therefore one can be tempted to conclude that the salivation becomes

a predictor for the reward "food". However, there are experiments which show that this interpretation is far too complex. It it has been observed that Pavlov's dog licks a light bulb when the light-bulb has been learned as the conditioned stimulus (Lieberman, 1993, p.354). The same applies to pigeons which learn that a light precedes the application of food (Domjan, 1998, p.62). The pigeons always *first* peck the light and then run to the food-dispenser (and peck the food). From that one can conclude that the stimulus food is simply *substituted* by the light and does not carry any higher semantic meaning in the sense of a reward. Especially in the case of the dog it is obvious that the licking the light-bulb is not very "rewarding" for the dog.

ISO learning also works with stimulus substitution. There is no interpretation of the primary stimulus (food or bump) as a reward or punishment. The reflex is simply substituted by a predicting behaviour. Therefore ISO learning leads to the same behaviour.

Next, the *flexibility* of the motor reaction has to be discussed. In all the above experiments the motor reaction has always been the *same*. Thus, the UR and CR were the same. This is also the case in ISO learning if only one output-neuron is considered. In that case ISO learning can not generate completely new reactions during learning since there is only one motor-reaction possible. The only parameters which can be varied are strength and timing. This limitation can be seen in avoidance reactions of rats. There are experiments which demonstrate that rats are able to escape an electric shock by jumping over a fence. However, the rats are not able to press a button to prevent the shock (Lieberman, 1993, p.354). More complex organisms, however, are able to generate different behavioural pattern to different stimuli. If one wants to model more sophisticated motor reactions one has to think about extensions of ISO learning at that point (Lieberman, 1993, p.168). The simplest extension has already been shown in the robot experiment where two neurons for speed and angle have been used. Already with this simple combination of two neurons the robot shows quite complex behaviour after learning.

5.4.2 Instrumental conditioning

In contrast to classical conditioning *instrumental conditioning*[5] explicitly uses a feedback loop in the environment: "if the occurrence of an operant is followed by presentation of a reinforcement stimulus the strength [of the operant] is increased" [cited from Skinner in Hilgard (1975)]. Thus, instrumental conditioning is a closed loop-paradigm since the behaviour (the operant) feeds back to a special input ("reinforcement stimulus"). An action (or a chain of actions) leads to a reinforcement of the action. Therefore in the context of feedback reinforcement-learning and ISO learning seem to be similar.

However, reinforcement learning explicitly involves a reward-signal. Therefore to describe ISO learning in the context of reinforcement-learning a reward has to be *defined*. On the level of observed behaviour this can be demonstrated by the robot experiment (see section 4.3). Assuming that there is no access to the internal structure of the organism the reward has to be defined by behavioural observations[6]. If the robot is observed performing its attraction task one could interpret finding (and "eating") the light-spots as a reward (and the bump into obstacles as punishment). Thus, it is possible to interpret the robot's behaviour in the context of reinforcement learning if one defines a certain behaviour as the reward. However, from observation it is difficult to say if the reward is the final turning reaction or the disappearing ("eating") of the light-spot.

Summarising, although there seems to be a reward in the attraction-simulation there is no reward-signal in the robot itself. Thus, the introduction of the reward does not relate to internal signals and is therefore an observer-problem which shall be avoided in this thesis.

---

[5]Instrumental conditioning can also be labelled with the expression "reinforcement-learning". Both terms will be used equivalently. In a strict sense there might be a difference between the two expressions, namely that reinforcement-learning can be interpreted as open-loop when it is used in a technical application where an engineer trains the network. However, the training is also performed on the outcome of an action and is therefore open-loop again.

[6]Such a behavioural observation can be anything which indicates a reward. This also involves self-observation like introspection. However, even self-observation can not observe neuronal *signals* but only mental states. Thus, there must be still a definition of a reward in the context of mental states.

5.4.3 How to distinguish between classical conditioning and instrumental conditioning?

Despite the problems of defining a reward properly it is possible to interpret the robot's behaviour in the context of reinforcement-learning and in terms of classical conditioning (with an appropriate feedback). Therefore the distinction which separates reinforcement learning from classical conditioning is much more fuzzy than expected. Psychologists are aware of this similarity for precisely the reasons which are pointed out above. Therefore they also conclude:

> Classical conditioning and reinforcement learning are much more the same than Skinner proposed (Hilgard, 1975, p.209).

Thus, the way to distinguish between classical conditioning and reinforcement learning by closed/open-loop or by reward/stimulus-substitution leads to unsatisfactory results. Finally it is an observer-problem since the observer can not be sure of having identified the right closed loop or having identified the right reward.

However, the distinction between classical conditioning and reinforcement-learning has not been given up. Therefore another distinction has been introduced: it is the number of sequential motor-reactions followed by a stimulus. Dayan (2001) argues that instrumental conditioning involves a *chain* of *multiple motor reactions* ("action planning") to optimise a *final* reward whereas classical conditioning involves only one final decision ("stimulus-response" or a "habit"). Therefore Dayan and others argue that there are two different systems which interact:

> Konorski, Dickinson, Balleine and their colleagues have suggested that there are really two separate motivational systems, one associated with Pavlovian motivation, as in SR[7], and one associated with instrumental action choice[8] (Dayan, 2001).

However, ISO learning still does not fit into this distinction. Looking at the circuit which represents the flow of the *signals*, ISO learning (see Fig 2.1) is clearly a "stimulus-response" (or classical conditioning) system since it directly transforms a stimulus into a motor reaction (this has been already pointed out above). However, looking at the *behaviour* of the robot in the attraction experiment the robot seems

[7]Means stimulus-response which is equivalent to classical conditioning.
[8]Equivalent to reinforcement-learning.

to perform a *sequence* of movements to find the target (namely the light-spot). Thus, it is a matter of the point of view (therefore of the observer) if ISO learning is interpreted as a stimulus-response system or as an action-planning system.

5.4.4 Classical conditioning and instrumental conditioning in the context of constructivism

In the above paragraph it was shown that it is possible to interpret ISO learning in the context of classical conditioning or in the context of reinforcement learning. However, in the context of constructivism certain aspects are not permitted and certain aspects are allowed.

In constructivism the concept of a "reward" is not permitted as being an attribution to an internal state of the organism. This would violate the separation of the system-levels. At this point it must be stressed that it is not the question if a reward has "really" a neuronal correlate or not. Constructivism simply avoids this dispute in not attempting to identify neuronal structures with the observed behaviour.

The introduction of the reward also leads to problems of how to interpret the actions which lead to the reward. This becomes clear if Dayan's definition of reinforcement-learning is taken which is based on "action-planning". Thus, the organism generates a sequence of actions to get the final reward. This leads to the assumption that organisms are rationally working towards a reward. The question arises: Do they "really" optimise their rewards? A reward optimising organism would explicitly plan its behaviour to get the final reward. Thus, first there is the plan and then there is the reward. However, one could argue that very often we behave the other way round. We stumble into a good outcome and then post-rationalise the actions which happened before as rational action-planning. Thus, maybe the career of a celebrity was simply a chain of lucky outcomes because he/she has been at the right places and met the right people. After having become famous the press, the PR-team or the celebrity him/herself invent stories regarding the sophisticated life-long plan to become famous[9].

All these problems arise when different system-levels are mixed up: A certain

[9]Some constructivists go so far to say that persons in general post-rationalise their life in inventing reasons why they have made certain decisions. Constructivists argue that persons can only perform self-observation since they are not able to access their internal neuronal states and are therefore in a similar position like an external observer.

behaviour is interpreted as a reward or a sequence of behaviour is treated as action planning. This might or might not be the case. Therefore an interpretation like Pavlov's stimulus-substitution avoids the endless debate of defining rewards in staying either on the neuronal or on the behavioural level.

Open loop or closed loop was another distinction to decide if learning is reinforcement learning or classical conditioning. In the context of constructivism only the closed loop models can be used. Reinforcement learning uses the closed loop model. However, it uses a reward-signal and is therefore not directly applicable. Classical conditioning is by definition open loop. However, with an appropriate feedback it can become a closed loop model without a reward (with stimulus-substitution) as discussed above.

Another important criterion in constructivism is that the organism shall control its input and not its output. This means that the organism acts in the environment to achieve a certain input-condition and not a certain output. This is the case in reinforcement-learning where the organism acts to get a reward when one accepts that the reward finally results from a certain input condition (which is usually not specified). Since classical conditioning is open-loop it can not sense consequences and therefore it can only have the task to control its output.

Summarising, neither classical conditioning nor reinforcement learning fits in the context of constructivism. The basic reason for this is the elimination of the reward by the system-levels in constructivism. Another reason is the closed loop character which demands that the organism has to control its input and not its output.

5.4.5 Models of animal learning: drive re-enforcement vs reward-based learning

After having introduced the learning paradigms from the psychological perspective they will be now be re-introduced from the perspective of computational neuroscience.

Temporal sequence learning has often been associated with classical conditioning (Pavlov, 1927; Dayan and Abbott, 2001). In classical conditioning an association between an unconditioned stimulus (US) and a conditioned stimulus (CS) is learned so as to learn a conditioned response (CR). Since *temporal* sequence learning learns the sequence between events one event must be the reference for all the

other events. Thus, a reference is needed with defines $t = 0$. Additionally a reference in form of a pathway or a signal is needed which drives the learning behaviour. Such a reference can be interpreted in two very different ways which leads to two very distinct groups of temporal sequence learning: the drive-reinforcement models (Sutton and Barto, 1981; Klopf, 1988) and the reward-based models (Sutton and Barto, 1982; Dayan, 2001).

In the drive-reinforcement models the strongest response triggered by a stimulus serves as a reference. Usually this is the unconditioned response (UR) triggered by the unconditioned stimulus (US) which results from a strong connection between the sensor input of the US and the motor output. Learning tries to generate an earlier conditioned response which anticipates the unconditioned response at $t = 0$. If the conditioned response has become strong enough it will replace the original UR; indeed it will now actually become a new UR on which further temporal sequence learning-stages could be built. Thus, learning is not guided by a pathway or signal with a special label, rather it is guided by the strength of the response called the *drive*. That learning is guided by drive reinforcement is supported by psychological studies and is called "stimulus substitution" (see above). ISO learning is clearly a drive-reinforcement model since it does not use any reward signal and since it is able to substitute one drive by another. Leaving all weights variable makes the ISO-algorithm completely un-supervised since there is no special input[10].

On the other hand there exists a variety of models which use a *reward*-signal as a reference and try to predict the reward in order to maximise it. Interestingly today only these models have survived. The development of temporal sequence learning-rules has completely shifted towards the reward-based models. By the use of the reward signal, these algorithms belong to the class of externally evaluating learning schemes. Learning-algorithms which need external evaluation usually have their applications in engineering where the (external) engineer teaches the system to make it useful for his/her purposes.

As a consequence they do not fit directly in the framework of autonomous behaviour in its rigorous sense. However, it is conceivable that reward-based learning-systems do exist in autonomous agents in the sense that they are bootstrapped by "first correlative experiences", for example by ISO learning.

[10]This is true for the learning circuit (Fig. 2.1). The special role of some inputs is not determined by the learning circuit but by the feedback loops.

5.4.5.1 The Rescorla/Wagner rule

Rescorla and Wagner (1972) were the first to try to describe classical conditioning in a formal mathematical model. Their aim was to explain the *development* of the response (UR,CR) in time, not its outcome. According to their theory, learning is driven by the surprise a stimulus represents for an organism. This surprise is measured in their model by the strength of the association between the US and the CS. The surprise is at its maximum before learning and converges to zero if the CR has the same strength than the UR. Therefore, the surprise can be measured as the difference between the maximum strength of conditioned response $V_{\max}$ and the strength of the conditioned response after trial n: V_n. The dynamics of the response V_n are described by the Rescorla/Wagner learning rule:

$$\Delta V_n = c \underbrace{(V_{\max} - V_n)}_{\text{surprise}} \tag{5.10}$$

At the beginning of learning the associative value between the CS and the US is zero and the surprise is maximal which leads to maximum learning. During the course of learning the surprise decreases due to the causal coupling of the CS with the US. At the end V_n has the same value as $V_{\max}$. The problem with the Rescorla/Wagner model is that the value for $V_{\max}$ is known only after the experiment has taken place. Thus, it is only possible to describe the learning dynamics a-posteriori and not in real-time during the experiment. However, the Rescorla/Wagner-rule makes it plausible that the surprise can be seen as a basis for learning. In our model this surprise plays an important role, too, but it is expressed in a completely different form, namely as the disturbance D in the closed loop model. Before learning the organism experiences the highest surprise (contingency) and after learning the organism is able to predict the disturbance. As a consequence the surprise has been changed to certainty.

5.4.5.2 The Sutton and Barto Model of classical conditioning

The Rescorla/Wagner-Model was not able to model classical conditioning in real-time since it needs the final outcome of the experiment in the form of $V_{\max}$. The earliest model which was able to model the process of ongoing-learning for classical conditioning was developed by Sutton and Barto (1981). It uses a similar learning rule to ISO learning involving the derivative of the output signal and correlating it with the input-signal (see Fig. 5.3a). However, there are important differences.

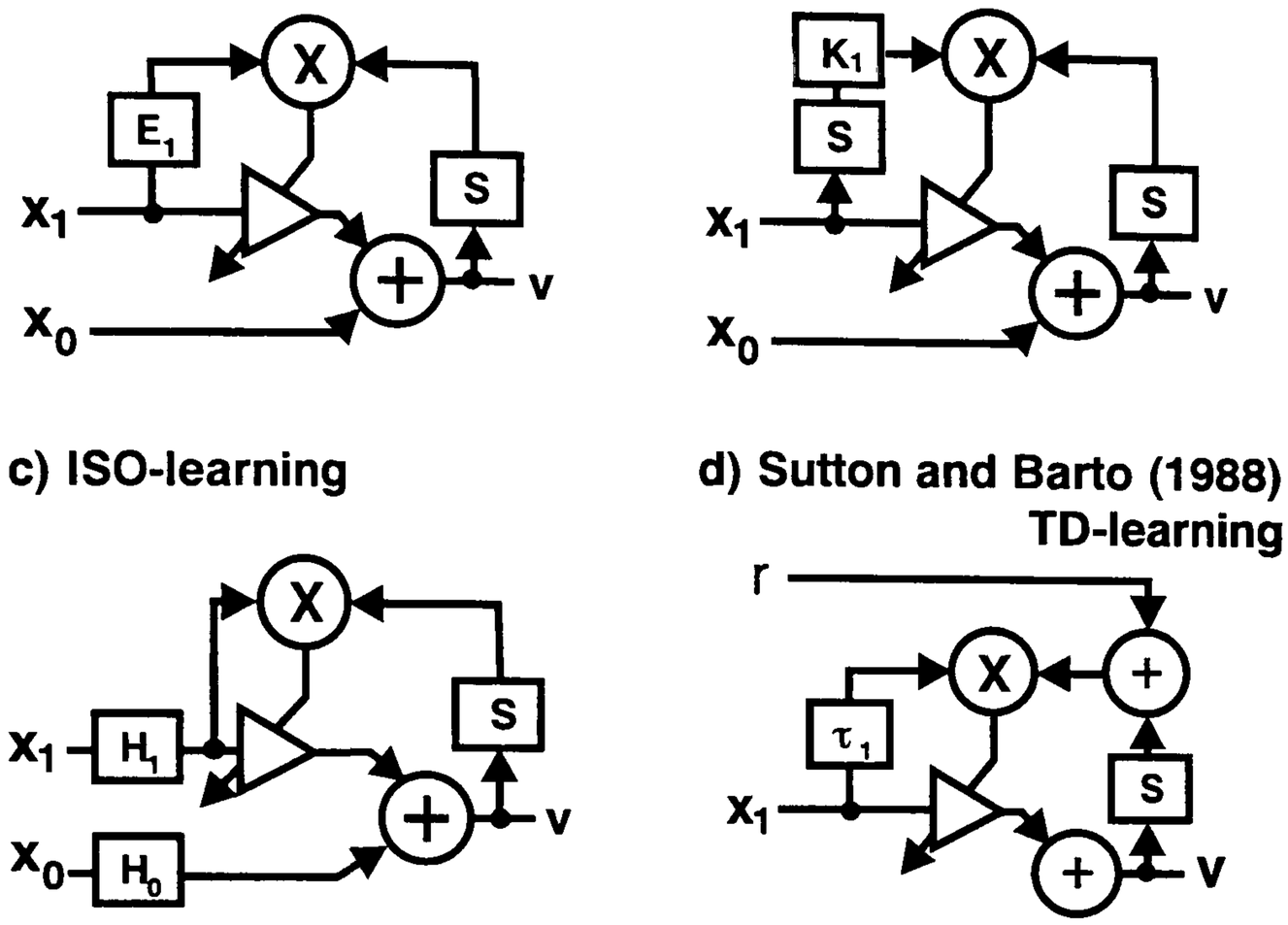

FIGURE 5.3: Comparison of three drive reinforcement algorithms (a-c) and TD-learning (d) in LAPLACE notation. Transfer functions are denoted as E, K, H, T, the derivative operator as s. The input X_0 represents the unconditioned (US) and X_1 the conditioned input (CS). All models are extendible to more than one CS but to reduce the complexity only one CS-input is shown. The amplifier symbol denotes the changing synaptic weight. Note that diagram (c) is drawn with a fixed weight at X_0 to make it more easily comparable to the other diagrams. All models use a derivative of the postsynaptic signal to control the weight change. Both Sutton and Barto-models (a,d) use low-pass filters K only in the conditioned pathway, Klopf's model (b) is identical to model (a) with the exception of an additional temporal derivative at this input. Only in ISO learning all inputs are filtered, which together with the output-derivative generates orthogonal behaviour, leading to weight stabilisation (for further explanations see text).

Each conditioned stimulus[11] $(x_k, k \geq 1)$ generates an "eligibility trace" in the form of an exponential decay which gives the system the opportunity to calculate temporal correlations. This trace[12] is calculated by

$$y_{k,t+1} = (1 - \lambda) z_{k,t} + \lambda y_{k,t} \tag{5.11}$$

[11]Not to confuse the reader with more additional symbols than really needed the naming convention of the thesis is taken. x still represents the input of a model-neuron and v represents the output of the neuron. All signals $x_k, k \geq 1$ are CS-inputs and the signal x_0 is the US.

[12]Note the similarity to the Kalman-filter theory and its signal model.

where $\lambda \in [0, 1]$ is a decay constant and

$$z_{k,t} = \begin{cases} 1, & x_k > 0 \\ 0, & x_k = 0 \end{cases} \tag{5.12}$$

This trace enters the learning rule

$$\Delta\rho_k = \beta\alpha_k(v_t - v_{t-1}) \cdot y_i \tag{5.13}$$

where α_k and β are learning constants. The output v of the system is calculated with the *unfiltered* inputs:

$$v = \rho_0 x_0 + \sum_{k=1}^{N} \rho_k x_k \tag{5.14}$$

First, the Sutton and Barto model is compared with ISO learning regarding the filters which have been used (E_1 and H_1 in Fig. 5.3a,c). The Sutton and Barton model uses a simpler filter at the input x_1 (or CS) than ISO learning. In contrast to ISO learning which uses filters of second order (see Eqs. 2.2 and 2.3) the Sutton and Barto model uses only a low-pass filter of *first order* (see Eq. 5.11). The decay constant a is always the same for all filters. This restricts the model to simple timing conditions which have been judged as being not realistically enough for classical conditioning experiments. ISO learning generalises from the Sutton and Barto model in using a filter bank of second order filters (band-passes) to deal with different temporal delays.

If one compares the structure of ISO learning with the Sutton and Barto model, it can be seen that there are different pathways for the processing of the input signals for the learning rule and for the generation of the output. This leads to the effect that the learning rule correlates *filtered* input signals (see Eq. 5.11) with the derivative of a sum of *unfiltered* signals (see Eqs. 5.13) and 5.14). In ISO learning the learning rule correlates *only filtered* signals with each other, namely *filtered* input signals with a sum of *filtered* signals. This symmetry of correlating *only* filtered signals with each other leads to the highly desirable feature of the stabilisation of the weights after the reflex has successfully been avoided.

However, the Sutton and Barto model does not aim for weight stabilisation depending on an *input*-condition ($x_0 = 0$) like in ISO learning. Their weight stabilisation is related to a certain *output*-condition and is therefore related to the model of Rescorla and Wagner (1972). In their model learning stops at the moment when

the CR has reached the same magnitude as the UR. Thus, in the Sutton and Barto model learning stops when the strengths of the UR and the CR are at the same levels. This model it is equivalent of having learned the relation between the CS and the US.

At this point it is quite obvious that the Sutton and Barto-model is an open-loop model since it implicitly assumes that the strength of the UR and the CR after learning should be the same. In the robot example it became clear that this is normally not the case which can be seen in the strongly different motor-responses before and after learning (see Fig. 4.3). The reflex behaviour can be quite coarse with strong reactions while learning can lead to more precise actions which need a small motor signal.

The Sutton and Barto-model presented here is a typical drive-reinforcement model since learning is driven by the strength of the signals and not by an external teaching signal. The model has not been developed further, since it failed to reproduce some psychological results properly as shown by Klopf (1988) who also provided an improved version of the Sutton and Barto model. Klopf's model shall be discussed next.

5.4.5.3 Klopf's model

The model by Klopf (1988) is shown in Fig. 5.3b (see also for a similar model: Kosco 1986). In particular the dependence of different intervals between the US and the CS on the learning rate has been improved. To achieve this Klopf basically used a more complex filtering of the input signals for the learning rule. While the Sutton and Barto model uses only a first-order low-pass with one parameter (λ in Eq. 5.11), Klopf uses an FIR-filter. An FIR-filter is implemented by a tapped delay line where every tap contributes a weight to the learning behaviour. Therefore the FIR filter has as many parameters as there are delay-elements. Thus, the FIR-filter offers much more freedom for the design of the learning behaviour than the IIR-filter in the Sutton and Barto-model as it has more free parameters. As in the Sutton and Barto model only the learning circuit (which changes the weight) gets the filtered CS-signal. The second difference to the Sutton and Barto-Model is the use of the derivative at the input of the learning circuit. Klopf argues that only *changes* in the CS-inputs cause change of the weights.

The actual weight change is calculated by the following equation which also incor-

porates the FIR-filter with the coefficients c_k.

$$\Delta\rho_{j,t} = \Delta v_t \sum_{k=1}^{\tau} c_k |\rho_{j,t}| \Delta x_{k,t-j} \tag{5.15}$$

The c_j are *fitted* to psychological data to mimic the effects of different timings between the CS and the US. The summation is the same as in the Sutton and Barton-Model (see Eq. 5.14) but every input signal x_i is split up connecting to one positive and one negative synapse to establish a more realistic model.

Klopf's model is also a drive-reinforcement model as it does not define an explicit reward. Because of the use of the derivatives and the application of the FIR-filter makes the model becomes robust against different temporal relations and durations between the US and CS. The Sutton and Barto model, for example, demands that there is no temporal overlap between the CS and the US while the Klopf-Model can cope with such an overlap.

With respect to this thesis, the most interesting aspect of Klopf's work is that he has taken the environment into account. He argued that the environment of an autonomous agent has to be non-evaluative. This means that it must not provide explicit evaluations, for example reward signals. All evaluations have to be performed implicitly within the organism's boundaries. Those implicit evaluations should be free of anthropomorphic interpretations and he argues that learning is only based on relating signals to signals:

> I will suggest that *drives* in their most general sense, are simply *signal levels* in the nervous system, and that *reinforcers*, in their most general sense, are simply *changes in signal levels* (Klopf, 1988).

This directly relates to the constructivist's view: The environment and the organism shall use descriptions which are free of attributions coming from any observers perspective. Taking the environment as non-evaluative gives one the opportunity to describe it purely by the laws of physics. The same applies to the organism itself, if one defines the drives and reinforcers only by the dynamics of signals.

Another interesting insight is given into positive and negative feedback loops. Klopf argues that any positive feedback must be combined with a negative feedback to ensure stability. Furthermore he argues that even positive feedbacks can be described as negative feedbacks:

> Drives implemented as positive feedback loops would seem to support the goal of drive induction rather than drive reduction. With this having been said, it may then be observed that, in the case of biological systems, drive induction, as in the pursuit of prey, always to be followed by drive reduction, as in the *consumption of prey* [emphasised]. This may suggest a simple general principle for the design (or evolution) of drive-reinforcement networks: primary drives implemented as positive feedback loops should always lead, when activated, to the subsequent activation of primary drives that are implemented as negative feedback loops (Klopf, 1988).

At that point one can go one step further and argue that in choosing the right sensor-motor loop it is possible to form a negative feedback also within the paradigm of food acquisition. This has been shown in the computer-simulation (section 4.3).

5.4.5.4 Temporal Difference (TD) Learning

Sutton and Barto developed their first model further which they called TD learning (Sutton, 1988). In contrast to their earlier model they introduced an explicit *reward* signal (see Fig. 5.3d). This signal represents an explicit goal in the learning algorithm which should be reached during learning, namely to predict the reward signal. TD-learning has the goal of generating an output v which predicts a reward r by the help of its (sensorial) input signals x. This goal is achieved by minimising a prediction error δ between reward and output. Thus, learning relies on the predefined reward which acts like a teacher signal in supervised learning (Widrow and Hoff, 1960). Another difference to the drive reinforcement models is that the output in the TD-model is no longer a motor-reaction, it is the prediction of the reward. Thus, with the introduction of TD-learning the output signal of the algorithm has become the status of an internal signal whereas the output of the drive-reinforcement models is interpreted as the conditioned or un-conditioned (motor-) response. The same applies to the reward signal itself. If one wants to describe an autonomous agent then he/she is forced to define what is a reward and has to hard-wire the reward(-system) into the organism. This carries the danger that the observer-perspective is imposed onto the system and at the end the organism is no longer autonomous but has become a slave of the external observer. This has to be kept in mind if one uses TD-learning to model autonomous behaviour.

Now the similarities between the TD-learning rule and the one used in the current study have to be discussed. The original TD-learning by Sutton and Barto uses discrete time steps (τ) and this shall be used as a basis here. However there is a time-continuous version available (Doya, 2000). TD-learning calculates a temporal difference error δ (thus, similar to the famous δ-rule by Widrow and Hoff 1960) by means of subtracting subsequent output values from each other and relating this error value to the reward:

$$\text{delta}_t = r_t + v_{t+1} - v_t \tag{5.16}$$

The actual weight change is then performed by correlating the result of Eq. 5.16 with the corresponding input-signal delayed by $n\tau$, with$n \geq 1$:

$$\Delta\rho_{k,t} = x_{k,t-\tau}\text{delta}_t \tag{5.17}$$

Note that in Fig. 5.3d only one CS-input is shown which enables TD-learning only to look one step ahead. In real applications TD-learning needs a tapped delay line for each CS-input which generates a sequence of CS-pulses (see Dayan and Abbott 2001 for a detailed description of TD-learning).

The second group of terms in Eq. 5.16 seems to be related to the derivative used in ISO learning (see Eq. 2.6). This mathematical similarity, however, carries a distinctively different interpretation, which can be understood as follows: The goal of TD-learning is that the output $v(t)$ should at any point in time predict the total remaining reward

$$v(t) = \sum_{s=t}^{T} r(s) \tag{5.18}$$

at the end of learning. Take the example of a rat exploring a maze where at each intersection a decision about a turn has to be made creating a temporal sequence of events. Each turn leads to a different reward (e.g., food) to be picked up along the way. This clarifies the concept of "total remaining reward" until the end of the maze is reached at T. Furthermore it is known that the total remaining reward can be iteratively approximated using the next following prediction value $v(t+1)$ to yield something like the total remaining *expected* reward:

$$\sum_{s=t}^{T} r(s) \approx r(t) + v(t+1) = e(t, t+1) \tag{5.19}$$

During learning this total remaining expected reward e is compared with its actual

prediction v to define the prediction error δ. Thus, delta$(t) = e(t, t+1) - v(t)$, leading to the apparent similarity of the resulting temporal difference terms $v(t+1) - v(t)$ in TD-learning with the derivative used by us. From this interpretation, however, it is quite clear that the term $v(t+1)$ arises only in conjunction with $r(t)$. This kind of conjunction cannot be found in ISO learning because it is reward-free. Furthermore, the structure of TD-learning is acausal, looking forward in time using $v(t+1)$ to calculate delta(t). In a strict sense looking into the future can only be performed by an observer who can *predict* the reward. Therefore it is not straightforward to implement the reward for TD-learning in an autonomous agent without violating causality.

The ideas of TD-learning are very similar to an algorithm used in engineering which is called "dynamic programming" which was introduced by Bellman (1957) and has been further developed under the name Q-learning (Watkins, 1989; J.C.H Watkins and Dayan, 1992). Bellman was interested in decision processes where during a multi-stage process the final outcome should be maximised:

$$x_N = t_N(t_{N-1}(t_{N-2}(t_{N-3} \ldots (x_0) \ldots))) \tag{5.20}$$

Here x_N should be maximised after having undergone the transformations $t_0 \ldots t_N$. A typical example is chess play where in every step a decision has to be made towards the final goal, namely to win the game. The idea is to solve this equation from the last transformation to the first one (Neuhauser, 1966). Therefore first the last transformation t_N is changed until a maximum has been reached and then the last but one and so on until t_0 is reached. This can be formulated by a recursion formula which is usually called the Bellman-recursion (Bellman, 1957, p.83). It is obvious that TD-learning and dynamic programming have several aspects in common, especially that both algorithms maximise the final outcome which is called reward in the case of TD-learning.

The direct comparison between ISO learning and TD-learning (see Fig. 5.4) shows that (as mentioned before) the reward pathway and the error calculation of TD learning is replaced by the reflex-pathway in ISO learning algorithm.

Both algorithms (TD and ISO learning) can be identified with neuronal structures. However, the structural differences of ISO learning and TD-learning suggest different neuronal substrates. The TD learning circuit consists of two different components: The error-signal circuit and the predictive circuit which are identified with the dopamine system and with cortical or other dopamine modulated brain areas. This is supported by the work of Schultz et al. (1997) who identified the response

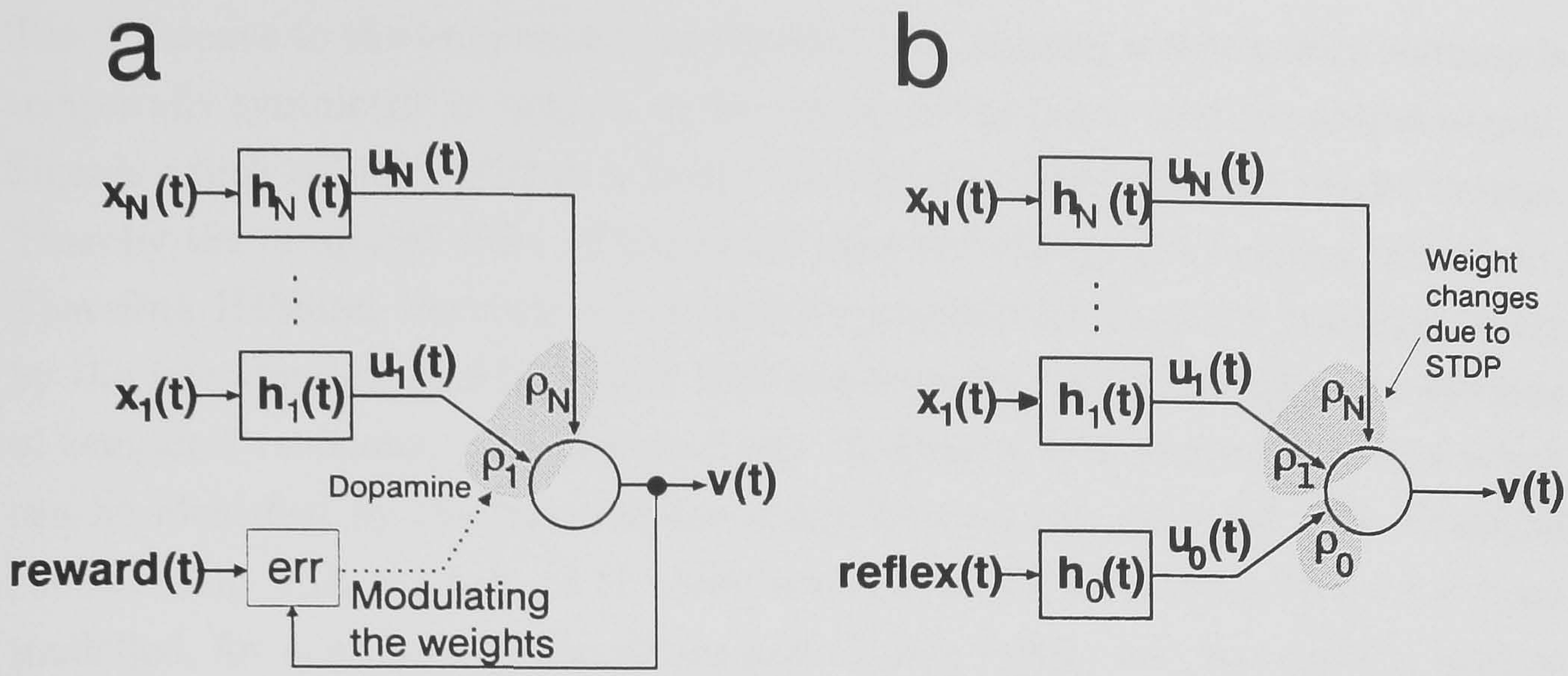

FIGURE 5.4: Differences between a) TD-learning and b) ISO learning.

of the dopamine-system with the error-signal from Eq. 5.16 in reward-experiments with awake monkeys. ISO learning algorithm, on the other hand, suggests only one neuronal circuit because all pathways are equivalent as supported by Hauber et al. (2001). They blocked NMDA-channels in instrumental conditioning tasks and could block the learning of the association between the reward and its predictors.

5.4.5.5 Motivated reinforcement-learning

At this point it must be mentioned that the concept of the reward has recently been extended to cover more psychological data. Therefore another consequent solution exists which instead of radically eliminating rewards (like constructivism) rather introduces rewards throughout. In Dayan (2001) classical conditioning is also interpreted as a reward-based theory so that it is possible to create a unifying theory built up on rewards which is called "motivated reinforcement-learning".

5.4.5.6 Pure Hebbian learning

Briefly it should be mentioned that it is also possible to use pure Hebbian learning to establish classical conditioning (Hebb, 1967). The Hebb rule is similar to many drive-reinforcement learning-rules (like ISO learning) in that sense that it correlates an input signal x_j with the output signal v and changes the weight ρ_j accordingly:

$$\frac{d\rho_j}{dt} = x_j \cdot v \tag{5.21}$$

The difference to the temporally *asymmetric* ISO learning is that such a learning is temporally *symmetric* in respect to the timing of the input- and the output-signal. Signals which coincide within a temporal window contribute to a weight change. Thereby the temporal order of the signal does not change the learning behaviour. Therefore Hebbian learning is not directly suitable for sequence learning. Only by the introduction of delays does Hebbian learning become suitable for learning of temporal relations. The simplest way to achieve this uses a delay line which can be identified by the transmission delay between two neuronal cells. Classical conditioning with the help of Hebbian learning and transmission delays has been modelled, for example, by Verschure and Coolen (1991) and successfully used in robot-experiments. Grossberg and Schmajuk (1989) used Hebbian learning in conjunction with a filter bank with different temporal delays to make it possible to learn different timings. Grossberg motivates the filter-bank approach by the fact that individual Purkinje-cells exhibit different delays so that a population of Purkinje cells can provide a filter bank which generates different delays. It is obvious that such a sequence of signals (Grossberg and Schmajuk (1989) call this a "spectrum") can be correlated with other signals using Hebbian learning. There have been several applications of this so called spectral timing model, for example, in pitch perception (Grossberg, 1995) or in motor control (Grossberg and Merrill, 1996). Also in technical applications the filter-bank approach has been used extensively, however mostly in the frequency-domain (Vaidyanathan, 1993).

The disadvantage of Hebbian learning is that the weights do not stabilise without additional measures being taken. If one wants to use Hebbian learning in the field of classical conditioning then the weights have to stabilise at the moment when the CR has the same magnitude as the UR. This can be achieved by delayed inhibition (Verschure and Pfeifer, 1992).

In Hebbian learning it is well known that the learning rule has a symmetric matrix which can be used to calculate its eigenvalues. The corresponding eigenvectors are the principle components of the input signals. It will be shown that such eigenvalues do not exist in ISO learning. The Laplace-representation of the learning rule Eq. 2.24 can be used to write it in a more general form and it is possible to describe both the classical Hebb-rule and the temporal Hebb-rule in the Laplace-domain:

$$M_{ij} = \frac{1}{2\pi}\int_{-\infty}^{\infty} U_i(i\omega)L(-i\omega)U_j(-i\omega)d\omega \tag{5.22}$$

$$M_{ji} = \frac{1}{2\pi}\int_{-\infty}^{\infty} U_i(-i\omega)L(i\omega)U_j(i\omega)d\omega \tag{5.23}$$

where $L(i\omega)$ determines if it is classical Hebbian or temporal learning

$$L = \begin{cases} 1 & \text{Classical Hebbian} \\ i\omega & \text{Temporal Hebbian} \end{cases} \tag{5.24}$$

In the case of classical Hebbian learning ($L = 1$) the change of the indices in Eqs. 5.22 and 5.23 does not change the result and therefore the resulting matrix is symmetric as expected. However, employing temporal Hebbian learning ($L = i\omega$) in Eqs. 5.22 and 5.23 changes the signs which makes the matrix anti-symmetric. As a consequence the matrix in the case of temporal Hebb has no eigenvalues. In general it can be stated that only in the case of a pure symmetric setup does the matrix have eigenvalues. That the matrix in the case of temporal learning has no eigenvalues reflects the property that classical Hebb learns events regardless of their temporal order while temporal Hebb learns events which form a temporal sequence.

5.4.6 Summary of the learning rules for animal learning

The last paragraphs were guided by the distinction between drive reinforcement models and those based an a value- (or reward-) system. What all these models have in common is that they analyse time sequences in order to generate anticipatory behaviour (for another summary of all learning rules except ISO learning see the technical report by Balkenius and Morén 1998). This is actually performed by analysing the time backwards starting at a certain reference-point ($t = 0$). This reference is either a reward signal (TD, Q-learning) or a drive (Sutton/Barto, Klopf, ISO learning). In dynamic programming and its successor Q-learning, the learning backward in time is performed explicitly in the form of decisions which are learned recursively. The rat, for example, decides at every branch if it should turn left or right to get the final reward. Once the final reward has been obtained the *last* decision is memorised and in the next trial the last but one decision will be learned. Thus, the rat maximises the final reward by starting with the final reward and then memorising the right decisions backwards.

However, the behaviour of a rat searching (and finally finding) food can also be explained by pure drive-reinforcement learning. In this case there is not a *final reward* but a *final behaviour* (namely the act of eating or the final movement towards the food). Predictions are related in the drive reinforcement models to the final behaviour. In the case of the reward-based models the eating of the food

is *interpreted* as a reward for the rat whereas in the case of the drive reenforcement models the eating of the food is *not* associated with any internal state of the rat. Thus, the drive reenforcement models do not attribute behaviour to internal states. They directly relate sensor inputs to motor reactions.

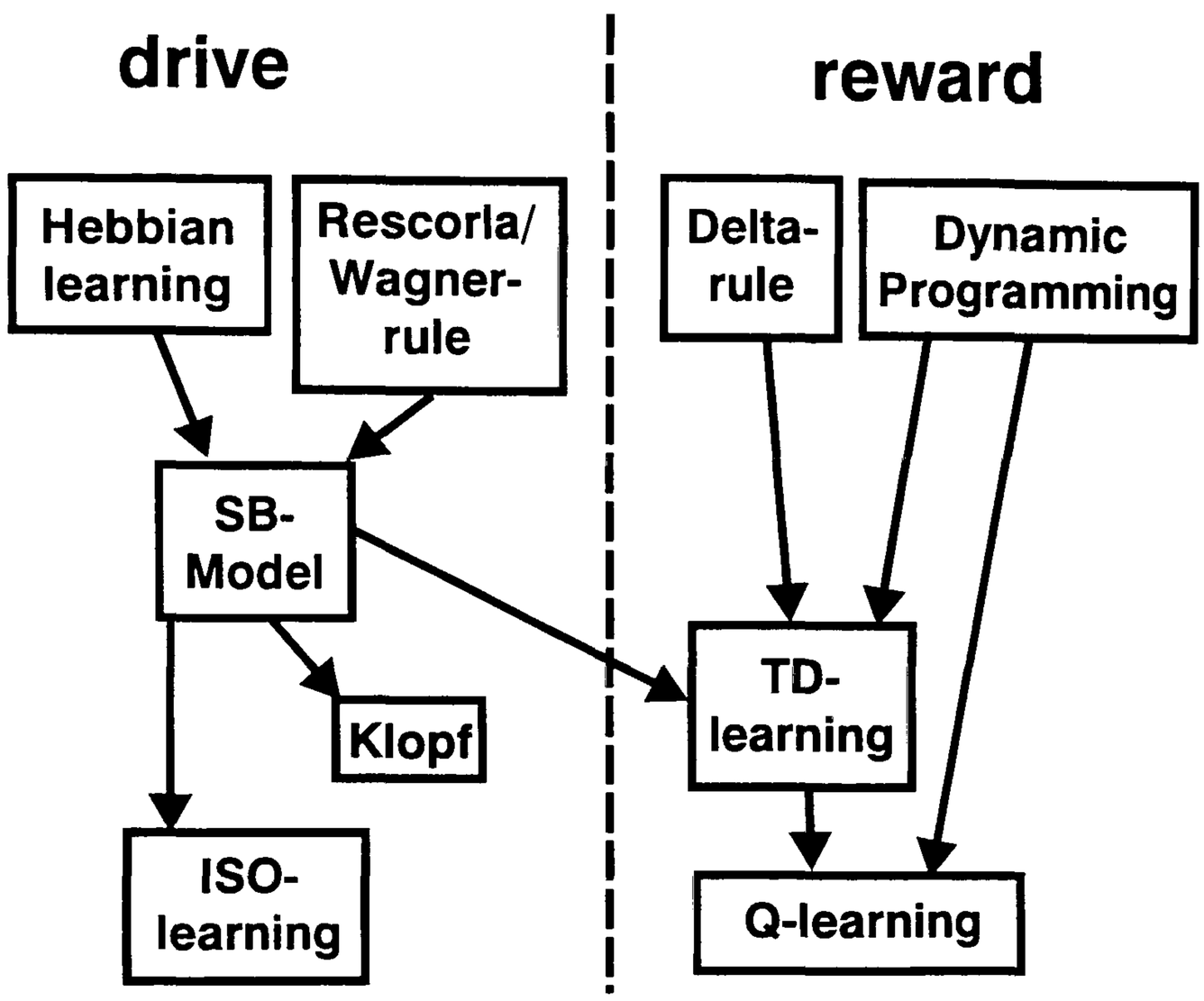

FIGURE 5.5: Comparison of the learning rules for temporal sequence learning. The rules can be related to psychology (observer perspective), to biology (neuronal perspective) or to engineering (tools). The actual implementations can be divided into drive-reenforcement models and reward-models.

A major difference between the different models is how or if the weights stabilise. In standard Hebbian learning the weights undergo exponential growth (Oja, 1982) so that the system deteriorates without additional measures. There are many solutions to solve the problem of exponential growth, such as well adjusted decay-terms (Young, 2001), synaptic competition (Miller, 1996b), restrictions imposed on the timing of the input signals (Klopf, 1988) and the already mentioned delayed inhibition (Verschure and Voegtlin, 1998)[13].

In the above mentioned models of classical conditioning (TD, Sutton and Barto, Klopf) stabilisation of the weights is achieved by taking the UR as a reference. The rules are adjusted in such a way that at the moment when the CR has the same amplitude as the UR learning stops and the weights stabilise. All these strategies to stabilise the weights share a common feature in that they use the *output signal*

[13] Also by personal communication at a conference in Edinburgh.

as a reference (in this case the UR). This dates back to the Rescorla/Wagner rule of classical conditioning where learning stops when the CR has a similar strength to the UR (see above).

In contrast to the above models ISO learning achieves weight stabilisation by its *feedback* which influences the *inputs*. From the moment the reflex has been avoided the weights stabilise. Thus, ISO learning is the only model which uses an *input* to stabilise the weights (see for example section 2.4.2 or section 2.5.2.3 where the weights stabilise when the input x_0 becomes silent). Using an input to control the learning behaviour only makes sense if there is a feedback from the output to the input so that the effect of the output on the environment and finally on the input can be felt. In the open loop case the weights continue to grow as in Hebbian learning (see for example Fig. 2.3 when x_0 and x_1 are active).

Another difference between the models is the level of biological or psychological realism. Q-learning is the most abstact form of sequence learning since it optimises a reward in a very formal way, like in a typical engineering task or in business-related optimising tasks. TD-learning, ISO learning, SB-learning and Klopf's learning rule claim to have a certain relation to biology in the sense that they are modelled with formal neurons.

5.5 Summary

This chapter compares ISO learning to open-loop paradigms in the fields of engineering, neurophysiology and animal learning. In the field of engineering the Kalman filter assumes low-pass filtered signals at its inputs because of their predictability. ISO learning goes along the same lines by making signals predictable at its inputs.

Recent results in neurophysiology have shown that the precise timing of pre and post-synaptic signals determines if a synapse is strengthened or weakened. The same applies to ISO learning where the timing of the input-signals determines if the weight is strengthened or weakened. The different parts of ISO learning can be partially identified by neuronal properties. The low pass filtering of the input signals can be identified with the passive low pass characteristics of the cell membranes and with the active properties of ion-channels, especially with the NMDA-channel. The important result is that with ISO learning the learning curve is directly obtained from the channel- and membrane-properties.

The main part of the chapter discusses the different learning schemes of animal learning and their corresponding mathematical models. The two main paradigms in animal learning are classical conditioning and instrumental conditioning. Looking closer at these two paradigms another distinction becomes more appropriate: learning with or without rewards.

The following discussion of the mathematical models also is guided by the difference between a reward being needed or not. ISO learning itself does not need a reward-signal and is therefore non-evaluative. Also non-evaluative are the early models by Sutton and Barto (1987) and the drive reinforcement-model by Klopf (1988). Although both models do not need any reward signal they are different from ISO learning. The difference arises in the different control strategies: The models by Sutton and Barto (1987) and Klopf (1988) control their outputs while ISO learning controls its input. This reflects the fact that the models by Sutton and Barto (1987) and Klopf (1988) are open-loop models and that ISO learning is designed for the closed loop.

Additionally, the reward-based model TD-learning by Sutton (1988) has been discussed. It is not directly related to ISO learning as it is evaluative. However, it looks similar due to its similar mathematical structure. Like ISO learning TD-learning utilises the derivative of its output signal. However, the derivative in TD-learning has another meaning than in ISO learning. In TD-learning the derivative helps to calculate the *expected reward*. ISO learning, however, calculates a motor output and not an internal signal, like a reward-prediction.

Chapter 6

Discussing the Organism in its Environment

6.1 Introduction

The last chapter discussed ISO learning without environment. In this chapter the environment is no longer ignored and establishes a closed loop condition. As in the previous chapters the closed loop is established by the environment by feeding the motor actions back to the sensors of the organism.

This chapter is divided into two main parts. The first part will discuss direct implications of the closed loop paradigm by comparing ISO learning to similar approaches, especially to approaches in the field of engineering. As described in this thesis, ISO learning tackles the problem of classical reactive control, in particular the fact that it always reacts too late. To overcome this problem ISO learning turns such a reactive system into a pro-active system. In the field of engineering similar problems arise when controlling a plant with a standard feedback-controller, for example with a classical PID controller. The solutions from the field of engineering and those by ISO learning will be compared in section 6.2.1.

The analytical treatment of ISO learning in the closed loop (in chapter 3) treated only two sensor inputs which means that only two loops can be created: the predefined reflex and the learned anticipatory (re)action. However, it is possible to employ more than *two inputs*. Consequently, the question arises if more than *two loops* are formed by using more than two sensor-inputs. This leads to nested loops and will be discussed in section 6.2.2.

The reflex pathway is the reference in ISO learning. The reflex pathway determines the direction of learning and defines what is early and what is late. Therefore the reflex pathway can be interpreted as a boundary condition for learning as pointed out in the introductory chapter. Section 6.2.3 will explore how other works in the field of machine- or animal-learning employ boundary conditions to prevent arbitrary results during and after learning.

The second part of this chapter will discuss indirect implications of the closed loop paradigm. As pointed out in the introductory chapter, the closed loop is the basis for an autonomous agent. This implies that the agent observes its environment in a different way than an observer observes the agent. One important implication of the observation process is observed uncertainty (section 6.3.1). The agent itself is confronted by the uncertainty of its environment. On the other hand, the observer is confronted with the uncertainty of the behaviour of the agent. Such uncertainty of behaviour could be interpreted as autonomy and therefore in section 6.3.2 an attempt will be taken to define autonomy by the observed uncertainty. However, in addition, the observer is usually an autonomous agent. When an observer observes uncertainty in an organism then the organism also observes uncertainty in the observer. This is called the "double contingency problem" and will be discussed in section 6.3.3.

There is also a conflict of interests between the organism and observer. While the organism itself wants to keep its homeostasis the external observer wants to treat the organism as an input-output system. Thinking of a hen and a farmer it becomes clear that the egg under the hen is observed in a different way. While the hen wants to keep the egg (homeostasis), the farmer wants to have the egg (input/output system). This example reflects the different perspectives of autonomous organisms and engineers. This will be discussed in section 6.3.4.

Finally the question will be asked if robotics can be used to clarify processes in biology: Is it possible to model neuronal processes on a robot or not (section 6.3.5)? The last section of this chapter will assume that this is in general possible and will discuss the different approaches in the field of autonomous robotics (section 6.3.6).

6.2 Anticipatory closed Loop Control

6.2.1 The Inverse Controller

In the robot experiment it has been demonstrated that the reflex implements a non-optimal solution. The bump can not be avoided by the simple reflex loop. However, by using predictive sensor inputs it was possible to generate anticipatory reactions so that the bump could be avoided. Since the sluggishness of the feedback loop is a very generic problem it does not only pose problems in biology but also in engineering.

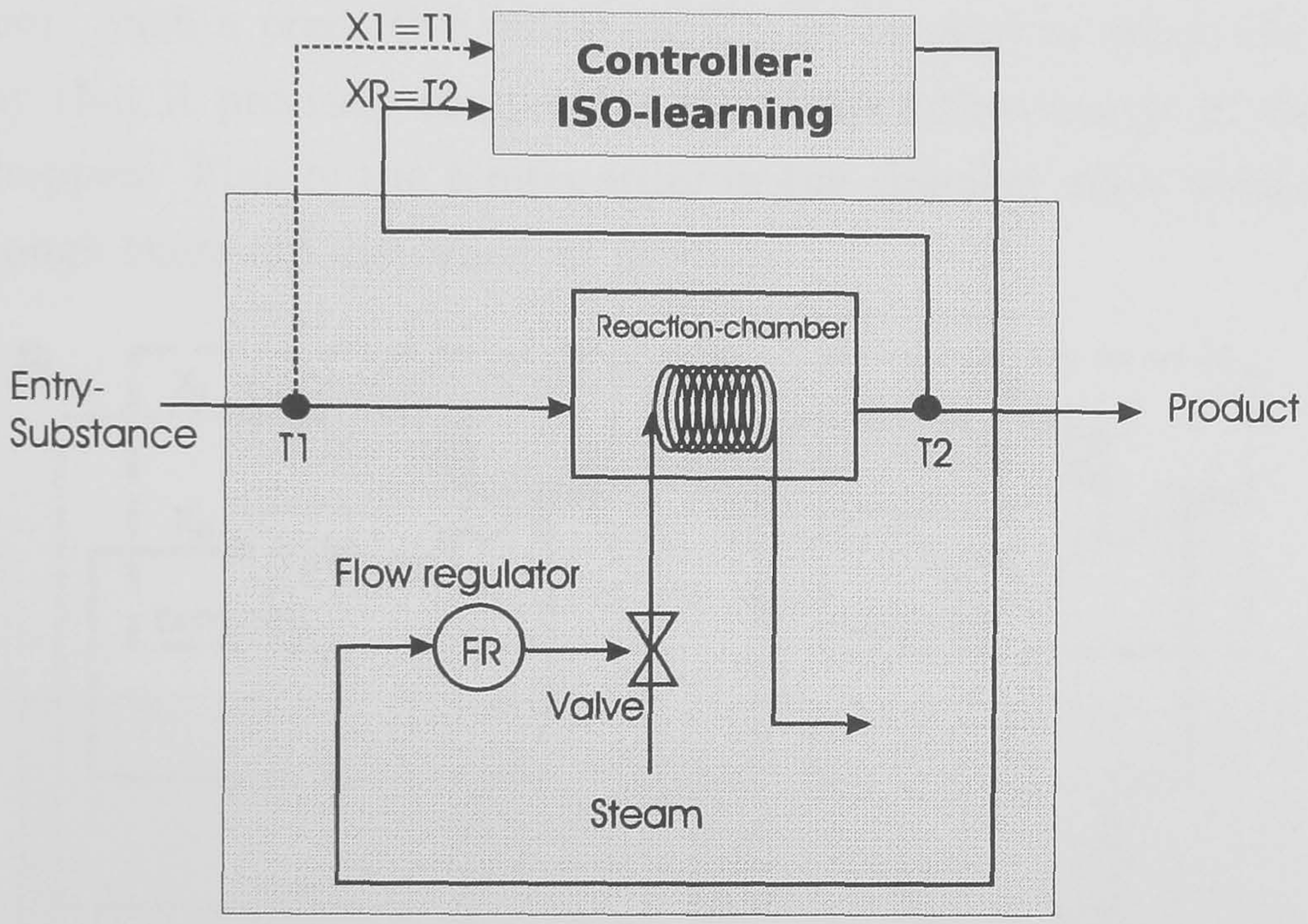

FIGURE 6.1: A possible application of ISO learning in a chemical plant. A reaction chamber transforms an entry substance with the help of heat into the final product. The heat is provided by steam and can be regulated by means of a valve. The chemical reaction of the chamber has an optimal temperature and therefore the task is to keep the temperature $T2 = const.$ This is achieved by a feedback mechanism involving $T2$, the controller and the valve which controls the amount of steam. The entry substance has the temperature $T1$ which can vary and therefore disturb the feedback loop. ISO learning can use the change at $T1$ (additional input as dashed line) to generate an anticipatory response.

As in biology the simplest form for an engineered control-process is the feedback loop (Palm, 2000). For example, with such a control the temperature in a reaction-chamber can be kept constant (see Fig. 6.1, solid lines). This is achieved by a closed loop involving a heater, a temperature sensor and an appropriate

controller. The controller generates from the temperature-signal an appropriate control signal for the heater so that the temperature is kept constant. However,

such a setup has the same problem as all feedback-controlled systems, namely that it only can react *after* the temperature in the chamber has changed. The unwanted temperature change is due to a disturbance in the environment which can not be predicted by the simple feedback controller. Thus, the temperature-change in this engineering example is the equivalent of a bump in the robot-experiment.

As in the robot experiment the feedback-controller which controls the reaction-chamber can be extended so that the unwanted sluggishness of the feedback can be eliminated. This is achieved by using sensor signals which are able to *predict* the temperature change in the reaction chamber. Such a predictor could be the temperature of the substance which is about to enter the chamber (see Fig. 6.1, dashed line). Such a predictive sensor-signal can be used to adjust the heater in such a way that it precisely counteracts the temperature-change at the moment it would happen. Finally the temperature in the chamber stays constant all the time, although there are disturbances present.

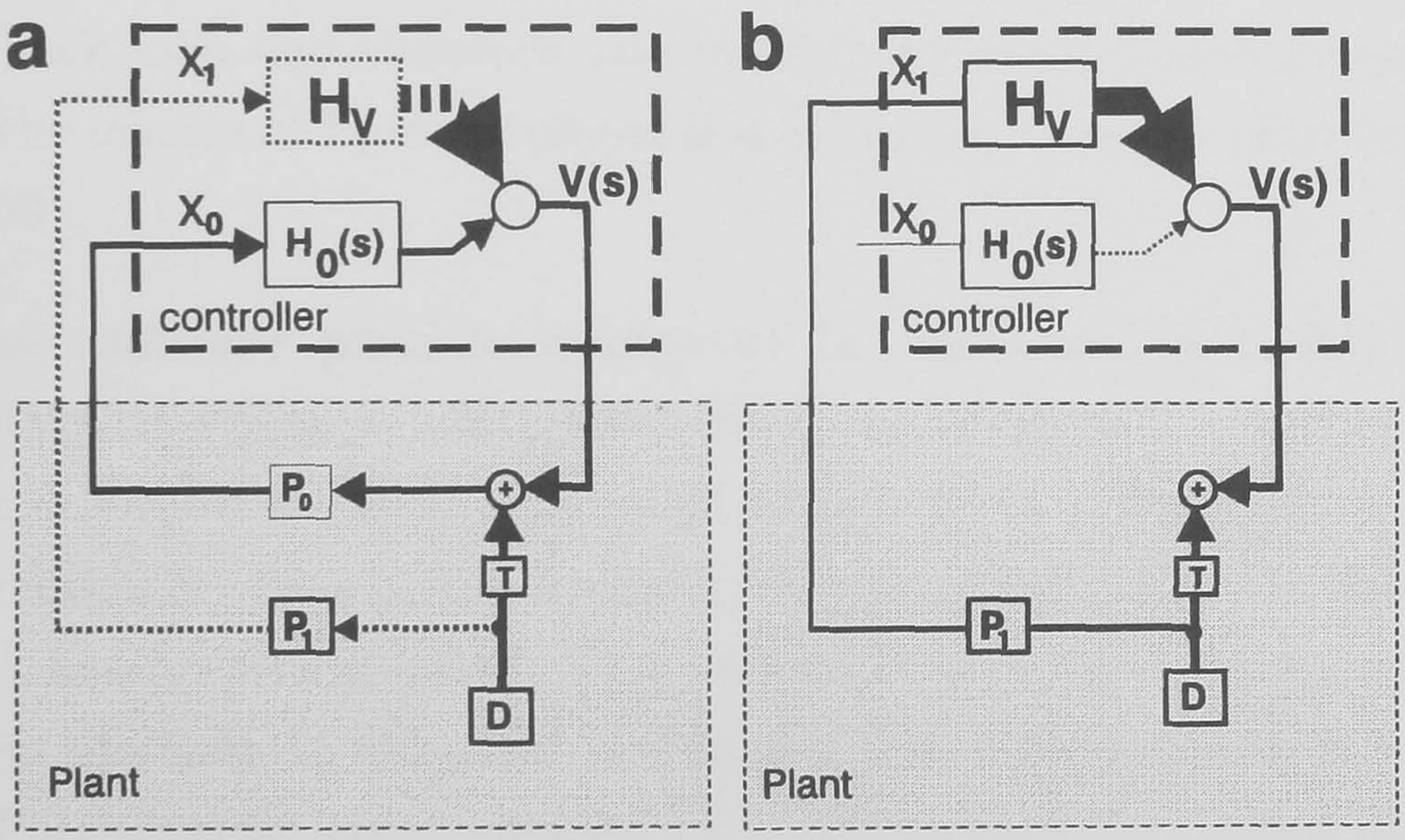

FIGURE 6.2: Illustration of the inverse-controller paradigm (based on Fig. 3.2 with $P_{01} := 0$). a) Controller before and during learning. b) After having successfully avoided the inner reflex loop (P_0, H_0, ρ_0).

Fig. 6.2 shows the generalised version derived from the above example with the reaction chamber. The transfer functions P_0 and H_0 form the feedback loop and P_1, H_V form the predictive pathway. The signal D is the disturbance. Fig. 6.2 is a simplified version of Fig. 3.2 where $P_{01} = 0$. Thus, the output of the controller does not affect the predictive pathway (D, P_1). In the reaction-chamber example this means that the heater does not change the temperature of the entry substance before it enters the reaction chamber. For example, the entry substance is stored in another place so that it cannot be affected by leaking heat from the

reaction chamber. Since the circuit-diagram is a special case of the derivations of section 3.3 the mathematical framework derived there can be directly applied to this engineering-problem.

Fig. 6.2b shows the condition where the controller is always able to keep the output at the desired state and therefore the feedback is no longer used. This is achieved by an appropriate H_V which generates with the help of the predictive input x_1 an output-signal so that the organism is able to counteract the disturbances. Having $P_{01} = 0$ Eq. 3.6 becomes

$$H_V = -P_1^{-1} e^{-sT} \tag{6.1}$$

Thus, the transfer-function of the controller H_V is composed by two transfer-functions of the plant. The delay T and the inverse of the transfer-function P_1. The difficult task for the controller is to find the inverse of P_1. A solution which approximates Eq. 3.6 by a superposition of resonator-responses has been shown in this thesis. However, this problem is far more general since it is present in every feedback loop and therefore also in every technical feedback-system. Since in Eq. 6.1 the inverse of P_1 is calculated it is called the "inverse controller problem" (Palm, 2000).

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 (Palm, 2000, p.592). Often the transfer function H_V is adjusted manually or heuristically until the feedback loop (H_0, P_0, ρ_0) has been eliminated. This technique is called "disturbance compensation". As opposed to this, ISO learning 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.

> A difficulty with disturbance compensation is that it is an open-loop technique in that it contains no self-correcting action (Palm, 2000, p.592).

Thus, control engineers always use their forward controllers 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 ISO learning. Fig. 3.2 clearly shows that the reflex will again take over if the predictive pathway fails.

A frequently addressed problem in biology is motor control (Kawato, 1999; Karniel, 2000; Doya et al., 2001; Wolpert et al., 2001) and especially the control of volun-

tary 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 ("efferent copy", "corollary discharge"). By now clear evidence exists for such a general mechanism. The details of how it is implemented, however, are still under debate.

The development of the inverse controller inside the boundaries of an organism can also be interpreted as a form of object recognition. However, one must be cautious since constructivism does not permit mixing the behavioural level with the signal-level. Therefore one has to find a closed description on the level of behaviour and on the level of signals. On the level of behaviour one can interpret the robot's behaviour as follows: Before learning the robot can only react after it has bumped into an obstacle. After learning the robot generates an anticipatory reaction (caused by the range-finders) before the bump happens. This could be interpreted by an observer as the robot having gained knowledge about the obstacle (in the sense that it will trigger the reflex). Therefore the observed behaviour, namely avoiding obstacles, could be interpreted as the *recognition* of the obstacles. The equivalent interpretation from the robot's perspective is the calculation of a forward-model which supersedes the reflex. Thus, the generation of a forward-model on the robot's signal-level could be interpreted as a form of object recognition on the behavioural level. In the experiments presented here the robot only learns to identify obstacles. Scheier and Lambrosios (1996) used such a form of sensor-motor learning to learn to *categorise* between different objects.

6.2.2 Nested loops

In the above section the inner feedback loop was replaced by a fast feed-forward pathway. In the moment the feedback loop has become inactive learning has reached its goal and therefore further learning is no longer needed. However, in an organism learning can often continue and can form new feedback loops.

There is an important difference between the engineering model (Fig. 6.2) and the biologically inspired model (Fig. 3.2) which enables the latter to continue with learning. Opposed to engineering P_{01} is usually not zero so that there exist two

nested loops in Fig. 3.2. The inner loop is the original reflex formed by H_0 and P_0. The outer loop is established by H_V, P_{01} and P_1. Once the inner feedback loop (H_0, P_0) has been eliminated the outer feedback loop takes over. Thus, a new feedback loop has been formed which also has the same disadvantage as any other feedback loop, that it always is too late. This provides the opportunity to continue learning which has again the goal to eliminate the outer(most) feedback loop.

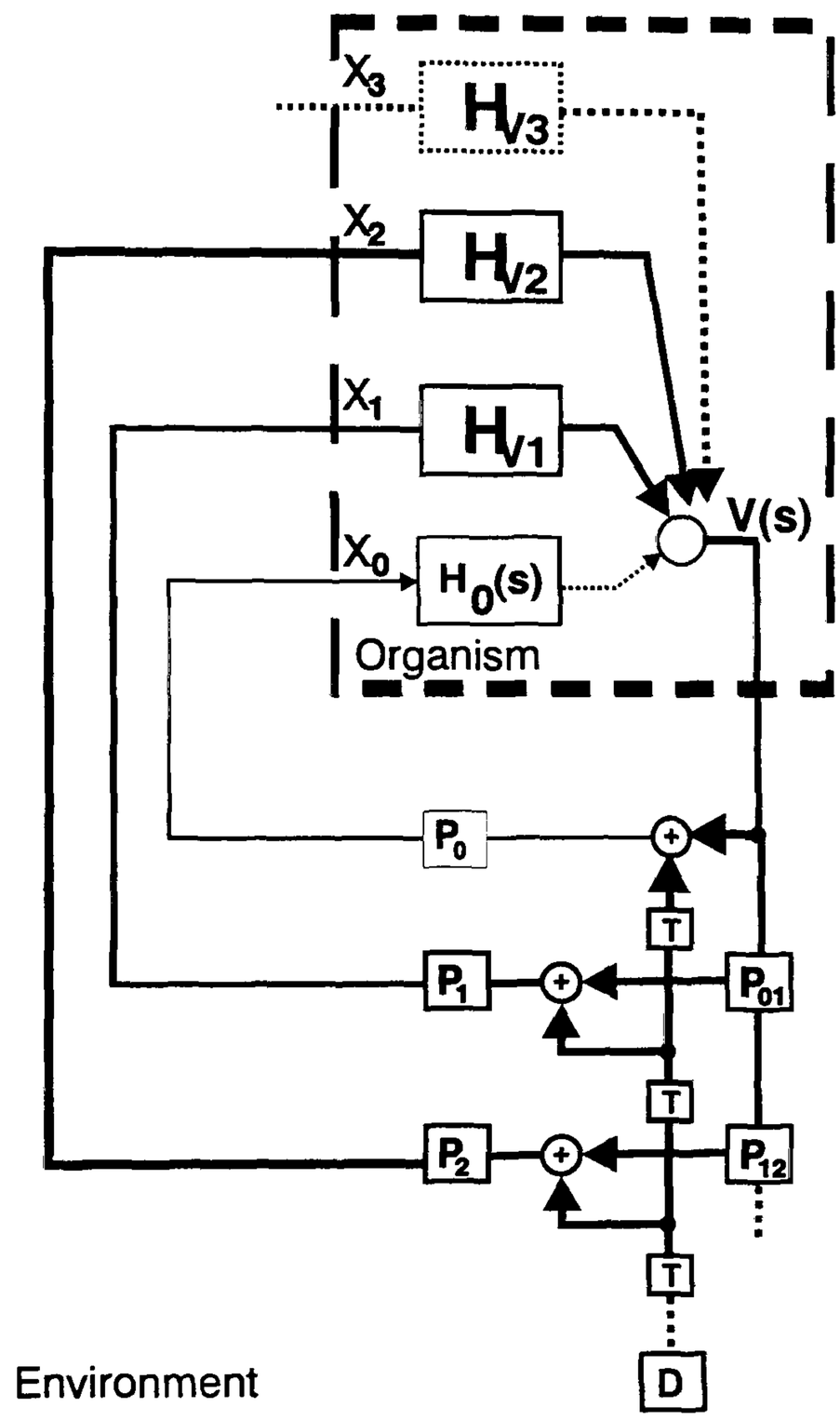

FIGURE 6.3: Nested loops

Fig. 6.3 shows a generalisation of Fig. 3.2 with more than two loops. Learning starts as usual by superseding the innermost feedback loop (H_0, P_0) by the second feedback loop (H_{V1}, P_{01}, P_1). This process can continue now over and over in superseding the mth feedback loop by the $(m+1)$th feedback loop. This process is limited by the number of sensor-inputs and by their ability to predict each other as there must be a delay between the different loops (T). Thus, *nested loops* are created which supersede each other finally leading to the loop which gets the disturbance without any delay. The concept of the nested loops has

several advantages. The fallback-principle is more gradual than in the case with only two loops. Once one of the loops fails, an inner loop can take over which gives learning more security. Also the use of more loops gives the organism the opportunity to develop a greater behavioural variety for different situations. For example in one environment the outermost loop works but in another situation only a loop in between works. Therefore the organism gets more flexibility.

Still the concept of loops has not been exploited to the end. Some year ago, von Uexküll (1926) argued that the sensor-motor-loops are only a part of the whole story. It is also quite obvious that the organism itself can establish *internal* loops within its boundaries. In ISO learning the next step towards an internal loop would be to use the motor output directly as an input without using the environment. Such a feedback mechanism is called efferent copy (von Uexküll, 1926; Grüsser, 1986) and is a consequent extension of the feedback mediated by the environment. However, from the organism's perspective there is no difference if the feedback is internal or external. If it is useful in the context of slow feedback loops it will be used.

6.2.3 Boundary conditions

Hebbian learning rules like the one used here belong to the class of unsupervised learning rules. Unsupervised learning seems to be the obvious choice for creating the first and earliest stages of autonomous behaviour, because it does not require external (teacher-like) knowledge. Instead it relies purely on self-organisation based on the correlation structure of the inputs. Such unguided self-organisation processes, however, can also lead to a situation where nonsensical correlations are learned leading in the end to an undesired network behaviour. The standard solution to avoid this problem is the introduction of boundary conditions which keep the self-organisation process within sensible margins. In practice this is either done heuristically by the network designer, or, as a better choice, boundary conditions are introduced such that they intrinsically (and in a natural way) represent the structure of the problem to which the self-organisation process is applied.

In the case of the unsupervised temporal sequence learning algorithm, this is achieved by embedding the learning circuit in an environment which leads to a closed loop situation. The causal relation which naturally exists between many different pairs of sensor events (pain follows heat, taste follow smell, etc.) as described in the introduction creates an implicit boundary condition for our al-

gorithm by using the latest incoming event (the one which drives the reflex) as the temporal reference for learning. The environment has two properties for ISO learning: It provides feedback and it contains disturbances, but it is clear that it does not provide any reward or any other teaching signal. Klopf (1988) called this feedback loop "non-evaluative" since there is nothing in the environment which evaluates the organism's performance. Instead, here ISO learning becomes *self-referenced* (von Foerster, 1960; Maturana and Varela, 1980): the actions of the learner influence its own learning without any evaluation process.

6.3 Observer-problems caused by the closed loop paradigm

6.3.1 Uncertainty vs. certainty

Coping with an uncertain environment is one main aspect of the definition of autonomous behaviour (Verschure, 1998). Ekdahl (2001) used the ability to *anticipate* events for his definition of autonomy. He distinguishes causal and acausal systems: a system relying only on reflexes is a causal system since it can not look into the future. Thus, the system experiences the environment as *uncertain* since it can never predict when the disturbance D will actually trigger the reflex-reaction (see Fig. 3.1). From the moment the system has built up anticipatory reactions the system has become acausal since it can to a limited degree predict the future. Thus, the disturbance D can be predicted and therefore the organism has gained *certainty* over the occurrence of the disturbance D (see Fig. 3.2). In this thesis words like "causal" or "acausal" must be used with caution since Ekdahl's definition of autonomy is in danger of getting mixed up with the definition of causal systems in the field of signal-/control-theory. In the context of signal-theory the robot's circuits always operate causally since the calculations of the signals can only be performed with signals from the past. However, for Ekdahl autonomous agents are those which are able to learn anticipatory behaviour. This distinction shall be used from now in this thesis and shall be extended by the observer-perspective. Thus, there shall be two different views: the organism's perspective and an observer who observes the organism. In the following the perspective of the organism is described first, followed by the observers perspective.

The acquisition of additional sensorial information enables the organism to predict changes in the environment. Thus, for the organism anticipatory actions with the

result from the prediction of a reflex reaction leads to more security as compared to a situation where it had to rely exclusively on the reflex *reaction*. However, on the other hand the gain of security for the organism will lead to an increase of uncertainty seen by any external observer. Or in other words: to an increase of uncertainty observed in the domain of behaviour. The uncertainty is expressed by the *behaviour* of the organism. This can clearly be seen in the robot-example: from the moment learning eliminates the stereotypical reflex the robot's behaviour becomes more unpredictable. The robot solves its goal (obstacle avoidance) but an external observer can only guess *how* the robot actually does that.

It must be noted that the robot still operates completely deterministically and that there is nothing mystical about that. However, the observer has the problem that he/she has no access to the internal structure of the robot (thus, it is a typical observer-problem). The more sensor inputs and the more nested loops exist the more the behaviour of the robot becomes unpredictable from the observer's point of view. At a certain point the observer is no longer able to conclude which sensor signal has caused a certain action. The observer has to begin to *guess* about the causes and consequences.

Thus, while the robot is gaining certainty about its environment the environment experiences the robot as uncertain. This duality of uncertainty and certainty depending on the point of view (organism vs. "observer") is often used in definitions of autonomy (Ford and Hayes, 1995). This will be explained next.

6.3.2 Autonomy

Based on the background of the two system levels (behaviour and neuronal signals) it now becomes possible to define autonomy. Autonomy shall be defined from the observer's point of view (thus, at the behavioural level):

> In colloquial speech, the more complex a system becomes, the more it hides its functioning and internal mechanisms from the curious observer, the more likely we ascribe purposeful behaviour to it. In man-made (i.e., allopoetic) machines designed so far, the purpose lies exclusively in the domain of descriptions of the observer (Riegler, 2002).

An organism is autonomous from the moment that the organism *shows behaviour* which is no longer completely predictable (Walter, 1953; Anderson, 1989; Riegler,

2002). This is always the case when the organism has more than one choice of what it could do but the observer does not know the cause of the organism's decision.

Finding out the reason behind a certain behaviour only is a problem for the observer since the organism itself is still, in theory, completely describable by its internal states (nervous signals, chemical potentials,...) but for us these internal states are usually not accessible.

Finally, it must be stressed that unpredictability can also be achieved without learning, because it is possible to design an agent which is hard-wired and which expresses random or stochastic behaviour (i.e. with an internal noise generator). Such a behaviour is completely unpredictable per se but not of interest in the context of this study.

However, there do exist theories which argue that at the beginning of the ontogenesis the organism is in a completely *unordered* state ("tabula rasa"). Learning has the task of structuring the organism step by step. Such a view is related to the so called synergetics of Haken (1992, 1995) and dates back to von Foerster's "order from chaos" (von Foerster, 1985). ISO learning and also, for example, Hawkins and Kandel (1984) oppose such a view that starts with (working!) reflexes and not with an internally unordered organism. Such an organism would probably not be able to survive. That an organism is already ordered at birth is supported by developmental theories, especially by Piaget (1930).

6.3.3 Double contingency

With more than one organism in the world each experiences each other as an additional source of disturbance and vice versa. Learning still has the task to make sure that every organism learns predictions about its environment. However, now "the others" are part of the environment and they also try to do the same namely predicting "their others". This leads to Luhmann's double contingency-theory: Mutual (viz "double") contingency is a basic phenomenon in which organisms try to predict each other. For Luhmann (1984, p.148) the double contingency is the driving force of any social system.

Double contingency already emerges when two organisms meet since they both will try to predict each other. This becomes clear in the robot experiment. If two robots were be placed in the playground they would bump into each other like any other obstacle in the playground. However, there is a difference between

walls as obstacles and robots as obstacles. Walls do not change their position so that the timing between the vision sensor and the bump sensor is completely defined by the approaching robot. Having two robots in the playground leads to the effect that the reciprocal anticipatory avoidance movements of both robots leads to a reciprocal change of the timing of the arriving sensor-events. Each robot now experiences difficulties in establishing temporal relations between the vision sensor and the bump sensor since the other robot will slow down when it predicts the bump with its opponent. Thus, reciprocal anticipations lead to much more complex learning than that observed with only fixed walls and objects. However, the aim of this section is only to give an idea of how a social system could emerge out of the duality of certainty and uncertainty. There is no attempt in this thesis to make it a serious topic since there are still too many open questions like how to measure observed uncertainty in the environment and how to relate it to the disturbance D without crossing the system-levels.

6.3.4 Differences between Biology and Engineering

Now, there shall be an attempt to demonstrate the differences between ISO learning and those of a typical engineering model (Luhmann, 1984, see footnote on p.63). In engineering there is always an external observer, the engineer, who wants the system (for example the robot) to do precisely what he/she bids. This can be achieved by hard-wiring all properties into the system or by "teaching" the system the desired response (Segre, 1988) which is also a standard technique in neuro-informatics (Pal and Kar, 1996). Before "learning" the neural network generates an undesired output or just generates a random-output. The engineer "teaches" the system by a special signal until it has reached the desired behaviour. The classical example is the delta rule where the actual output of a model-neuron is compared with the desired output. The error between the actual output and the desired output changes the synaptic weights with the goal to get the error to zero (Widrow and Hoff, 1960). Thus, the system first exhibits unpredictable or undesired behaviour. Then later (after learning) it becomes completely predictable in the sense that it is now *useful* for the engineer (who is part of the environment). ISO learning embedded in the environment behaves the other way round: for an observer, the behaviour of the robot at first is completely predictable due to its reflex. After learning the robot's behaviour is only partially predictable for an observer since the robot has found one behavioural solution out of many possible solutions. From experiment to experiment (and even during an experiment)

the robot develops different strategies so that, despite the fact that the robot always starts from the same pre-wired initial condition (reflex)[1]. This is the *complete opposite* of a technical solution: in a technical solution the observer wants to have a predictable system. Thus, one can differentiate between two different paradigms: the "Engineering Paradigm" and the "Biology Paradigm". The "Engineering Paradigm" is always interested in a particular desired behaviour which is achieved by an external evaluation of the system's behaviour. In the "Biology paradigm" the organism follows its internal objectives and there is no external evaluation.

Von Uexküll used this difference between biology and engineering and argued that machines can never be alive since they are only extensions of our sensor and motor surfaces. Obviously von Uexküll referred in his work to the engineering-paradigm which leads to reliable machines which indeed can be used to extend our senses (TV, radio, telescopes, ...) and our motor reactions (car, air-plane, ...). The biology-paradigm, however, leads to very unreliable machines since these machines become autonomous. Nobody would use these machines for his or her purposes since they produce uncertainty for their environment, hence for their users, too. Thus, it is clear that one has to choose the paradigm depending on the research-interest. Using the "biology paradigm" makes it possible to get closer to an understanding of autonomous agents.

Temporal difference (TD-) learning and also Q-learning have their origins in Engineering. This becomes clear when it is remembered that both algorithms use a reward-signal. In the context of engineering this makes sense since the engineer wants to have a reliable response at the end of learning. If one wants to use TD-learning rigorously in the context of autonomous behaviour ("biology-paradigm") then one is faced with a problem that the reward signal would have to come from "inside the organism" and not from the outside in the form of "wishful attributions" (Sharkey and Ziemke, 1997). Drive reinforcement learning eliminates the problem of defining a reward from the beginning and should be considered if an autonomous agent has to be designed.

[1]The pre-wired initial condition could be interpreted as a "genetic" basis.

6.3.5 Biology and Pure Physics

Descartes was one of the first who struggled with the problem that an organism should in principle be entirely describable by physical laws (Descartes, 1952)[2]. This poses the problem that there is no difference between an organism and a mechanical machine. The concept of a "soul" was no longer needed. Knowing that the church would never accept a view which explained a human only by physical laws he solved this problem by dividing the human brain processes into voluntary and in-voluntary parts (Rachlin, 1976, p.4). The in-voluntary processes in the human body could be explained by mechanical or physical laws whereas the voluntary processes could not.

Three hundred years later this discussion is still vividly alive: Von Uexküll distinguishes between biology and physics (von Uexküll, 1926, p.71) arguing that biology is more than only physics. While physics only relies on physical laws biology has an underlying "plan". This can be interpreted in different ways (God, metaphysics, ...). The "plan" in Uexküll's view means that the reflexes of the organism are perfectly integrated into the environment. This means that before learning starts there is already a perfectly adjusted mechanism, namely through its reflex. This perfect integration is due to evolutionary processes that the species has undergone during several millions of years. This thesis demands the same, however by bypassing evolution and adjusting the *reflex* so that "it works". In that sense this thesis completely conforms with Uexküll. However, there are different views regarding whether a robot can ever be perfectly integrated in the environment which leads to the next section.

6.3.6 Robotics

Robotics is a discipline which can clarify the concepts of autonomous behaviour and interactions with a complex environment quite naturally. The emphasis shall be on those contributors to that field who explicitly or implicitly use a closed loop paradigm.

Rodney A. Brooks is one of the pioneers in the field of autonomous robots (Brooks, 1989b; Lorigo et al., 1997). Brooks argues that looking for the right representation of the world for the robot is the major obstacle in designing a working robot and that the search for the right representation is endless. In the end the robot's

[2]Descartes: 1596–1650

representation is in danger of being a a representation which mainly reflects the engineer's world-view. Therefore Brooks make a radical decision and introduced robots without "representation". Instead he started from a *functional* point of view: His robots should "*work*" in their environment. Brooks used, for the definition of a "working" organism, Uexküll's suggestion that an organism has to be integrated into its environment where it shall always be able to perform its *function(s)*. Therefore for Brooks it is not important how the internal circuits of the robot are interpreted by an observer but how they perform a certain function while interacting with the environment[3].

If there is more than one sensor-motor loop these loops are organised in a subsumption architecture which finds its correspondence in this thesis in the nested reflex loops:

> We build an incremental layer of intelligence which operates in parallel to the first system. It is pasted on to the existing debugged system and tested again in the real world. This new layer might directly access the sensors and run a different algorithm on the delivered data. The first-level autonomous system continues to run in parallel, and unaware of the existence of the second level (Brooks, 1997).

Such engineered robots exhibit complex behaviour and observers attribute rewards, punishments and other anthropomorphic aspects into the robots. Brooks sees himself as an engineer and not as an psychologist or a biologist. Therefore he always refrained from implementing his robots in a biologically realistic manner. Brooks is of the opinion that the brain is still not understood and that the crude simulations done by connectionists are far removed from the realism which is needed to simulate certain brain structures successfully. Therefore Brooks operates on the level of *behaviour*. This point of view conforms with the behaviourists but in Brook's case also with the constructivists. Speaking with Luhmann he describes his robots only on the level of behaviour (see introduction) and does not make any attempt to cross levels. Thus, he does not attribute from behaviour to internal states and therefore he does not need a representation of "fear" or "pleasure" in the circuits of his robots.

[3]The discovery of the functional nature of neuronal processes especially in the retina by Lettvin et al. (1959) has probably stimulated Maturana to develop his theory of constructivism. The frog's retina is only interested in small moving objects (flies) and in big moving objects (enemies). The resulting motor-reactions are quite obvious and lead to two independent closed loops. The one has the function for eating food and the other loop has the function to escape.

Moving on to the signal (or neuronal) level Brooks identifies two ways to interpret the internal signals of his robots without using external attributions. One way refers to the functional cycles by von Uexküll as stated above. The other interpretation is related to constructivism. Constructivists argue that the neuronal system operates self-referentially since it relates neuronal signals to neuronal signals. As Brooks is an engineer he relates electrical signals to electrical signals:

> An alternative decomposition makes no distinction between peripheral systems, such as vision, and central systems. Rather the fundamental slicing up of an intelligent system in the orthogonal direction dividing it into *activity* producing subsystems. Each activity, or behaviour producing system individually connects sensing to action. [...] Our favourite example [...] is a creature, actually a mobile robot, which avoids hitting things. [...] It is still necessary to build up this system by decomposing it into parts, but there need to be no clear distinction between a "perception subsystem", a "central system" and an "action system" (Brooks, 1997).

This is closely related to Klopf's work and also to the approach presented here. ISO learning fits perfectly with Brook's view since it also only relates signals to signals. As in Brook's work this thesis does not use the term "representation". Instead *transfer functions* are used here which also simply relate signals to signals. Just as in this thesis Brooks is also aware of the observer problem. Attributing "fear" to a mobile robot does not mean that the robot actually has signals which represent "fear".

More biological realism towards biology can be seen in Verschure's work with mobile robots. In the field of temporal sequence learning Verschure has been working several years in using robot applications (Verschure and Pfeifer, 1992; Verschure and Voegtlin, 1998). In his words every organism undergoes three steps of development (Verschure, 1998): pre-wired reflex (fixed connections), adaptive control (classical Hebbian learning of sequences of sensor inputs) and reflective, contextual control (goal- or reward-oriented learning). In Vershure's terminology adaptive control has no goals but builds up temporal associations with "proximal" and "distal" sensors. At the stage of the reflective control a goal is introduced in the form of a reward or punishment when, for example, an object has successfully been found. Other have introduced similar distinctions between different levels of processing such as Meysel (1991) or Karniel (2000).

TABLE 6.1: Different forms of embodiment.

Authors	Embodiment	Organism	Environment
Searle/Sharkey/Ziemke	organismic	living	real
Brooks	functional	physical,living/artif	real
Riegler/Quick	self-referent	any	any
Pfeifer/Scheier	physical	physical,living/artif	real

refers to a physical object which determines with its boundary what is inside and what is outside. The question arises if a physical body is needed to implement an autonomous (or intelligent) agent.

Classical AI denies that an agent needs a physical body. For classical AI the (physical) agent is not important since the agent is completely describable by its underlying *algorithms*. Algorithms have the advantage that they are not linked to a special body or substrate (Dorffner, 1991, p.7). Therefore the actual implementation of the algorithm does not matter. This means for an agent that it can be either implemented as a computer-program and or as a living agent made of flesh and blood (Turing, 1950). Computers and living organisms are *equivalent* in classical AI. Both perform information-processing: they receive input-data, processes the data and produce an output. A summary of classical AI can be found in Pfeifer and Scheier (1999, pp.36–58).

The identification of an organism as disembodied information-processing computer by classical AI has always been criticised. All these criticisms target the disembodied view of AI and claim that a real body is needed. Therefore the subject which discusses whether a real body is needed for an intelligent autonomous agent or not is called "embodiment". The remainder of the paragraph will present different definitions of embodiment (see Table 6.1). It will start with the strongest definition of a "organismic" embodiment and close with "physical" embodiment.

The earliest criticisms which challenged classical AI came from Searle (1980). Searle compares a real organism of flesh and blood with a replica-model which is built in the form of a computer. The difference for Searle is that only the real organism can "really" feel "fear" or "punishment". In other words: the living organism gives the sensorial stimuli a *meaning* which can be further related to intentionality (Mele, 1997). Intentions are related to internal motivations which give the stimuli meanings. For Searle it is clear that only "real" organisms can have "real" feelings and that artificial agents do not experience feelings at all. Such *internal* representations only exist in a living organism and not in a mechanical

"body" (which is called today "organismic embodiment" Ziemke 2001). He argues that the existence of internal human-like experiences is linked to the special matter living organisms are constituted.

Searle's definition about "organismic embodiment" can also be found today:

> Without an integral body it [the robot] does not experience pleasure or pain in reinforcement learning; there are only weight changes or program changes. The actual putative 'experience' of a robot undergoing reinforcement learning is the same both for reward and punishment. The organism, on the other hand, is driven by its bodily aversions and needs (Sharkey and Ziemke, 1997).

Only if there are living cells and only if they form an integral body it is possible for the organism to have "real" experiences. "Real" feelings for Ziemke and Sharkey are linked to the matter they emerge from and how this matter is organised:

> A robot is a collection of inanimate mechanisms and non-moving parts that form a loosely integrated physical entity. [...] By way of example, if you attach a hula-hoop of a bunch of clothes pegs to your body, they will clearly be objects attached to your body. [...] There is not the same clear distinction between the robot body and the objects around it as there is for an organism. This is not just a trivial matter. The chemical, mechanical, and integrating mechanisms of the living things are missing from robots. Cells need oxygen and so living bodies need to breathe, they need nutrition and so bodies need to behave in a way that enables ingestion of appropriate nutrients (Sharkey and Ziemke, 1997).

However, Sharkey and Ziemke do not justify their view by referring to intentionality. They argue that *evolution* is the key to the difference between artificial agents and living organisms. Citing von Uexküll they argue that the organism's relation to its environment has developed during a long evolution and that finally a perfect solidarity between the environment and the organism has evolved. This solidarity is further developed during the ontogenesis of the organism. Thus, the organism is "rooted" in its environment. More specifically the organism has developed during its evolution "working" feedback loops (also called functional cycles) which are further developed during the ontogenesis.

At that point Sharkey's and Ziemke's arguments can be integrated into the framework of this thesis. In their view feedback loops have evolved in a long evolutionary process so that they perform their task for the organism. Sharkey's and Ziemke's point is that only a very long evolution is able to adjust the parameters of feedback-mechanisms. Otherwise the organism would deteriorate. This is contradicted by classical control-theory. Feedback-mechanisms are used because they are very robust and already a rough adjustment of the parameters let them achieve their goals. This can be seen in the robot-experiment. The retraction-mechanism which follows after a bump can have a wide range of parameters. It does not matter how the robot performs the retraction-mechanism. The important aspect is that the *result* is the right one (it must "work"). Therefore the argument that evolution has carefully adjusted the parameters of a feedback-mechanism does not necessarily hold. The advantage of feedback-mechanisms is that they know *very little* about the environment – but they still work. This also means that the robot is far more robust to *other* environments than predicted by Sharkey and Ziemke. Also the avoidance case makes it clear. First, the avoidance case has been simulated on a *computer* which provided ISO learning with an *artifical* environment. Later the same algorithm was transferred with the same parameters and was connected with a real robot. There was no need no change the parameters of the feedback loops. The real robot worked well with the parameters obtained from the simulation. Obviously, the difference between simulation and the real-world application is compensated by the feedback loop. Therefore this suggests that feedback loops provide a robustness which enables an organism to live with only roughly adjusted parameters and therefore a "perfect solidarity" between organism and environment is not needed. This perfect solidarity might later develop during learning (or the ontogenesis) but for initial reflexes it is not needed.

Above it has been stressed by Sharkey, Ziemke and von Uexküll that organisms operate in *feedback loops*. However, classical AI interprets an organism as an *input/output* system by ignoring the feedback loop. This is a direct implication of the computer-metaphor where a computer (or algorithm) is perceived by its user as an input/output system. The user inputs information, the computer processes it and sends it back to its output. There is nothing wrong if this metaphor is applied to computers as they have to be reliable tools. However, computers do operate in a closed loop as the user closes the loop by him- or herself. In a more systems theoretical interpretation the difference between input/output paradigm and the closed loop paradigm lie in the control condition. While the organism controls its input the computer user controls the output of the computer.

Brooks was probably the first who targeted the input/output paradigm in the context of autonomous agents (Brooks, 1989a,b). He argues that an organism, in contrast to an algorithm, is not an input/output or stimulus/response-system but rather a closed loop system. Like Uexküll Brooks stresses the fact that agents evolve in functional cycles (or feedback loops).

The concentration on functional cycles avoids attributions towards internal states. Brooks stays either on the signal-level or on the behavioural level but does not mix them. Therefore he calls his view "intelligence without representation". The agents *act* in an "intelligent" way but the internal wiring is guided by the demand that the feedback loops have to work. Brooks therefore avoids the never ending discussion if there are "real" feelings and if these are only represented in "real" bodies[4]. Thus, Brooks avoids the observer-problem, namely identifying internal states like feelings in an organism or a robot. As with this thesis he simply avoids giving an answer to the question by concentrating on the *function* of the agent. If the observer attributes pleasure or pain into the robot is his/her fault (and many do).

It is important to note that such functional cycles imply an environment. Otherwise there is nothing to do. In the case of Brooks the environment must be a real one to call an agent embodied. The agent itself can be made of something different than of flesh and blood.

> The problem of *embodiment* refers to the fact that abstract algorithms do not interact with the real world. Rodney Brooks forcefully argued that intelligence requires a body (from Pfeifer and Scheier 1999).

Therefore Brook's definition of embodiment is weaker than the one by Ziemke and Sharkey. Also an artificial agent can be embodied (see Tab. 6.1).

The most general definitions have been proposed by Riegler (2002) and (Quick and Dautenhahn, 1999) who demand that an embodied system must operate self-referentially and it must maintain internal goals (see Tab. 6.1). This is usually called autopoiesis. Quick stresses the point that the organism has to be structurally coupled to an environment and the environment must provide perturbations (see Maturana and Varela 1980). Therefore Quick emphasises, like Brooks, that there must be an environment and this must be different in contrast to the organism. Riegler on the other hand stresses the functional cycles (e.g., the feedback loops), namely that the organism has to "work" in its environment and that

[4]He also avoids the symbol-grounding problem which is discussed in Pfeifer and Scheier (1999).

it has to gain functional *competence* in its environment. This can also be called "historical embodiment". Riegler is probably the only one who stresses the fact that *learning* (gaining competence) is an important aspect of embodiment.

Both views are strongly related to this thesis where also self-reference is demanded and the disturbance from the environment plays an important part. These two points shall now be discussed in more detail.

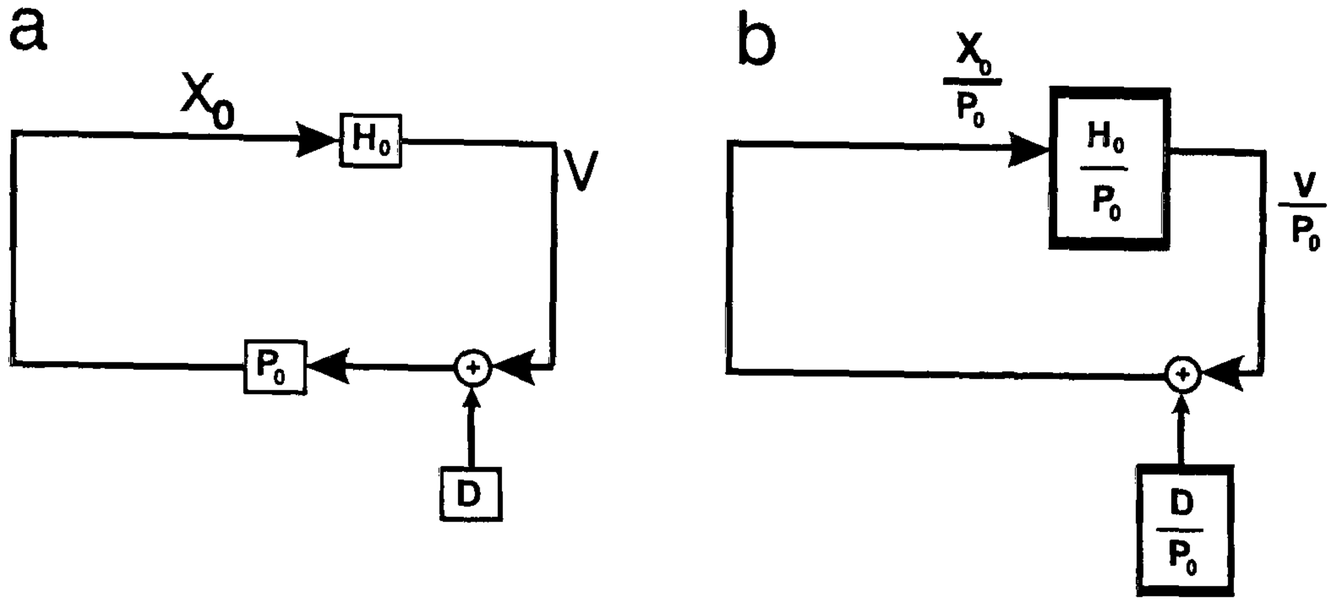

FIGURE 6.4: Transformation of the standard feedback loop (a, see also Fig. 3.1) into a unity gain feedback (b). The transfer-function of the environment P_0 can be integrated in the transfer-function of the organism. The environmental transfer-function can be eliminated so that the environment in the form of P_0 is no longer existing. However, the disturbance in the environment can not be eliminated.

Self-reference with a disturbance can be established by the simple reflex. This is shown in Fig. 6.4a. The question arises why these two demands lead to a definition of embodiment. More specifically it boils down to the question: what belongs to the environment and what belongs to the organism?

To find the answer to what belongs to the organism and what belongs to the environment it must be recalled that Fig. 6.4 represents the *organism's* point of view. The signals are therefore *neuronal signals*. Also the environment is represented as neuronal signals and therefore there is no distinction between environment and organism in the form of different signals.

Even the distinction between environment and organism with the help of the transfer-functions is not useful since Fig. 6.4a can be transformed to Fig. 6.4b by dividing by P_0. This effectively eliminates the transfer-function of the environment. In engineering this flexibility is often used to simplify the mathematical description of a system (Palm, 2000). The resulting transfer-function H_0/P_0 seems now to be inside the organism. Therefore from the organism's point of view it is difficult to decide what is inside and what is outside.

However, the *disturbance* can not be eliminated. It is always a property of the *environment* which can therefore be used to identify the environment. Therefore embodiment from the organism's point of view can be defined by the disturbances:

> A self-referential system is embodied if there exist disturbances which only exist *outside* the organism. Inside there are no disturbances. As a consequence a boundary which distinguishes inside and outside can be drawn.

Thinking in terms of evolution the disturbance can be interpreted as *the* constituent aspect for the organism. For example, cell-membranes have been developed during evolution in order to protect proteins from the contingencies (for example acids) in the environment. Therefore embodiment can be seen as a form of boundary-maintenance (Luhmann, 1984). This principle is used in Luhmann's work also on the level of behaviour. Society forms sub-systems which is also a form of boundary-maintenance (political parties, the financial system, gangs, ...).

With the above given definition of embodiment and the introduction of transfer-functions for the organism and the environment, a solution for the "hoola hoop" problem can be offered: does it belong to the "body" of the robot or not? Throughout this thesis signals have been related to signals by transfer-functions. Thus, if the hoola-hoop does not change any transfer-function it does not exist for the robot/organism. If it changes a transfer function it is inevitably relevant for the robot. If the hoola-hoop belongs to the organism or to its environment is a matter of interpretation. From the robot's point of view it is not distinguishable. As pointed out above the environmental transfer-function P_0 can be integrated into the internal transfer-function H_0. Using the above definition the hoola-hoop belongs to the robot's body if it does not cause a disturbance (for the robot). For example, perhaps the robot can exploit the dynamics of the hoola-hoop to perform a certain task better than without it. Thus, the hoola hoop is in this case is no longer a disturbance but it is integrated in the self-referential processes of the robot. This becomes clearer with an example: On the WGW'02[5] Pfeifer described a person who has to carry water in buckets down a hill. This person used the dynamical properties of the buckets filled with water to "dance" down the hill. Pfeifer added that the belly of the water-carrier might also contribute to the dynamics of the "dance". The question arises what belongs to the body and what does not belong to the body. Therefore he suggested "fuzzy" boundaries

[5]EPSRC/BBSRC International Workshop Biologically-Inspired Robotics: The Legacy of W. Grey Walter 14-16 August 2002, HP Bristol Labs, UK

between body and environment. This is equivalent to the possibility of changing the transfer-functions in the above mentioned manner (for example, of having only one internal transfer-function H_0/P_0).

In one example the hoola-hoop changed the robot's transfer-function(s) in a desired manner. Therefore, it was considered being a part of the robot's body. This argument can be made stronger. Pfeifer and Scheier (1999) demands that the agent has to be a *physical object*. Therefore this form of embodiment is called "physical embodiment". This means that the environmental transfer-functions (P_0 and P_1 in Fig. 3.2) have to be at least partially constituted by a contact to the physical world.

Summarising, like Riegler's and Quick's work this thesis does not demand that an autopoietic system has to consist of flesh and blood, nor does it demand a real environment. The consequence of this is that embodiment has nothing to do with the actual physical realisation of the agent. It can be an organism, a robot or a computer simulation. Any system which establishes autopoiesis and experiences perturbations can be declared as being embodied. Autopoiesis and perturbations translate in this thesis to feedback and disturbances which means that there is an active process which maintains homeostasis. Thus, embodiment in this thesis means more than a passive exchange between a system and its environment. This process has to be *active* and therefore a granite outblock in the antartic tundra (Quick and Dautenhahn, 1999) is not embodied from the point of view of this thesis. Active feedback is seen here as *the* basic property of the living since it implies boundary-maintenance (system/environment) and it directly provides a learning-goal, namely to supercede the feedback by fast feed-forward action. As already pointed out also the disturbance is essential for the definition of embodiment given here since it defines an area (the body) where this disturbance is compensated.

In a broader context embodiment can be interpreted as the creation of a boundary which has the task of reducing *entropy* within boundaries and is therefore related to the second law in thermodynamics (Balian, 1991). As discussed in this thesis, time plays an important role in this law since entropy is a measure of the unpredictability of events. Therefore this thesis argues that the basic driving force for the development of spatial boundaries and the development of suitable learning-rules is the a-symmetry of time.

6.4 Summary

This chapter has discussed ISO learning in the closed loop established by the environment. In the closed loop situation ISO learning turns the reactive organism into a proactive organism. A similar problem is known in the field of industrial control. Standard feedback control reacts also always too late. The solution of this problem is the inverse controller which performs feed-forward control. This is equivalent in ISO learning with the generation of an anticipatory action. Therefore ISO learning can also be applied to industrial control problems.

ISO learning does not limit the number of input channels. Consequently the number of feedback loops is not limited. Every input in ISO learning can therefore form a new feedback loop. New loops can be formed as long as new inputs are available and as long as the new loop anticipates the slower reacting loop. At the end, nested loops arise which anticipate each other. Such nested loops were also employed by Rodney Brooks in his subsumption architecture. In contrast to Brooks in ISO learning the loops emerge while in Brook's work they were usually hard-wired into the robot.

The closed loop paradigm has also consequences on the way the organism observes its environment and how the organism is observed by its environment. Especially, uncertainty is observed in a different way from the perspective of the organism and from the perspective of an external observer. While the organism gains security by learning anticipations, the environment experiences the opposite. The behaviour of the organism becomes more and more unpredictable. Consequently, autonomy is defined by gaining more certainty from the perspective of the organism and at the same time becoming more unpredictable for observers in the environment.

Not only the uncertainty is observed in a different way. Also the function of the agent is observed in a different way by the organism itself and by the environment. While for the organism its function is defined by itself, the observer defines the organism as an input/output system. The organism has to be useful to the observer. These two points of view have been called the "engineering paradigm" and "the biology paradigm". Consequently, one has to decide which paradigm should be employed in a certain context.

Autonomous robotics is the natural discipline which employs closed loop applications. In the context of this work it is important that robotics in particular is aware of problems which arise when behaviour is observed and interpreted. Rodney Brooks has shown that "intelligent" robots can be implemented without

attributing towards internal states. His solution is similar to the one presented in this thesis: closed loops. As already pointed out in the first chapter, closed loops establish systems without any quality. By closing the loop one can either stay on the level of signals or on the level of behaviour. This prevents the mixing of the system-levels and prevents misinterpretations of behaviour. This should also be kept in mind when working with reward signals in autonomous robotics. Reward signals cross the system-levels since the internal reward-signals are associated to external behaviour.

Chapter 7

Concluding remarks

Using constructivism as the underlying paradigm it was possible to develop a *reward-free*, isotropic algorithm for sequence order learning (ISO learning) in which learning relies only on the temporal order of its inputs. This has the advantage that all input signals are treated equally and that learning takes place between all of them. Thus, it represents a form of *unsupervised* sequence learning. Learning is only driven by the temporal relation between input- and output-signals.

In the second part of this study a closed loop situation has been introduced by means of behavioural feedback which determines the functional role of the inputs to ISO learning. The starting point is the setup of a primary reflex loop which is distinguished from all other inputs only by the fact that it initially carries the largest synaptic weight. In general, such closed loop reflex loop situations have the disadvantage that any *re*-action will only occur *after* an incoming sensor event.

This inherent disadvantage of feedback loops leads to a general objective for improving animal behaviour which is to find a mechanism which prevents the reflex. Sequence learning can achieve this by creating earlier, anticipatory actions.

In addition, it has been shown that weights stabilise as soon as the reflex has been successfully avoided. Because of the isotropy of the inputs, any other input line can take on the role of the reference signal during learning and the initial reflex can even be unlearned or reduced in strength – a situation which is observed in many physiological reflexes.

In the robot application it has been shown that ISO learning can solve the classical obstacle avoidance task in a fast and robust way. The robot was initially equipped with a fixed reflex reaction. ISO learning established then a relation between the

trigger of the reflex and earlier arriving signals from range finders. This lead to an avoidance reaction which prevented collisions with obstacles.

It seems that only avoidance-behaviour can be learned as ISO learning is guided by "reflex-avoidance". However, attraction behaviour can also be learned by ISO learning without any modification of the earning rule. Only the reflex must be adjusted. This has been shown in a simulated robot-experiment. This experiment made clear that one must be cautious when behaviour is interpreted and conclusions are drawn towards internal states. The behaviour of the robot suggested a reward-based maximisation inside the robot. However, internally it was a reward-free minimisation, namely the elimination of the reflex. To avoid such observer-problems this thesis suggests the sole use of one self-referential system-level: either neuronal signals or behaviour.

Appendix A

Plancherel's theorem

This theorem is rather unknown, therefore we state it here as:

$$\int_0^\infty f_1(t) f_2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F_1(i\omega) F_2(-i\omega) d\omega \tag{A.1}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} F_1(-i\omega) F_2(i\omega) d\omega \tag{A.2}$$

where F is the Laplace transform of f (Stewart, 1960). If we set $f_1 = f_2 = f$ it becomes the more commonly used theorem of PARSEVAL.

Appendix B

The robot-hardware

A modified commercial robot ("rug warrior", 16 cm diameter) was used. Two active wheels are driven by DC motors, steering is achieved through different DC-levels. Average speed was adjusted to 0.45 m/s using an appropriate bias to ds. To detect mechanical contact the robot has three microswitches in a triangular configuration. Visual signals are generated by active range finders with an angle of 70° between them. The computations were done on a computer (Pentium 90) running LINUX in realtime-mode. The communication between the robot and the computer was achieved by a simple cable.

Fig. B.1 shows the circuit which connects the computer (a) with the modified robot (b). On the robot-side only the additional components compared to the original design of the rug-warrior are shown. However, only the range-finder circuitry of the original robot was used.

The analog signals were provided by a cheap ISA AD/DA-card ("super 12 bit AD/DA-card"). Only the DA converter was used in the robot experiment. The A/D converter could be used to transmit the information from the LDRs for future experiments.

All *digital* signals were interfaced by the parallel printer-port. See table B.1 for the pinouts of the printer-port.

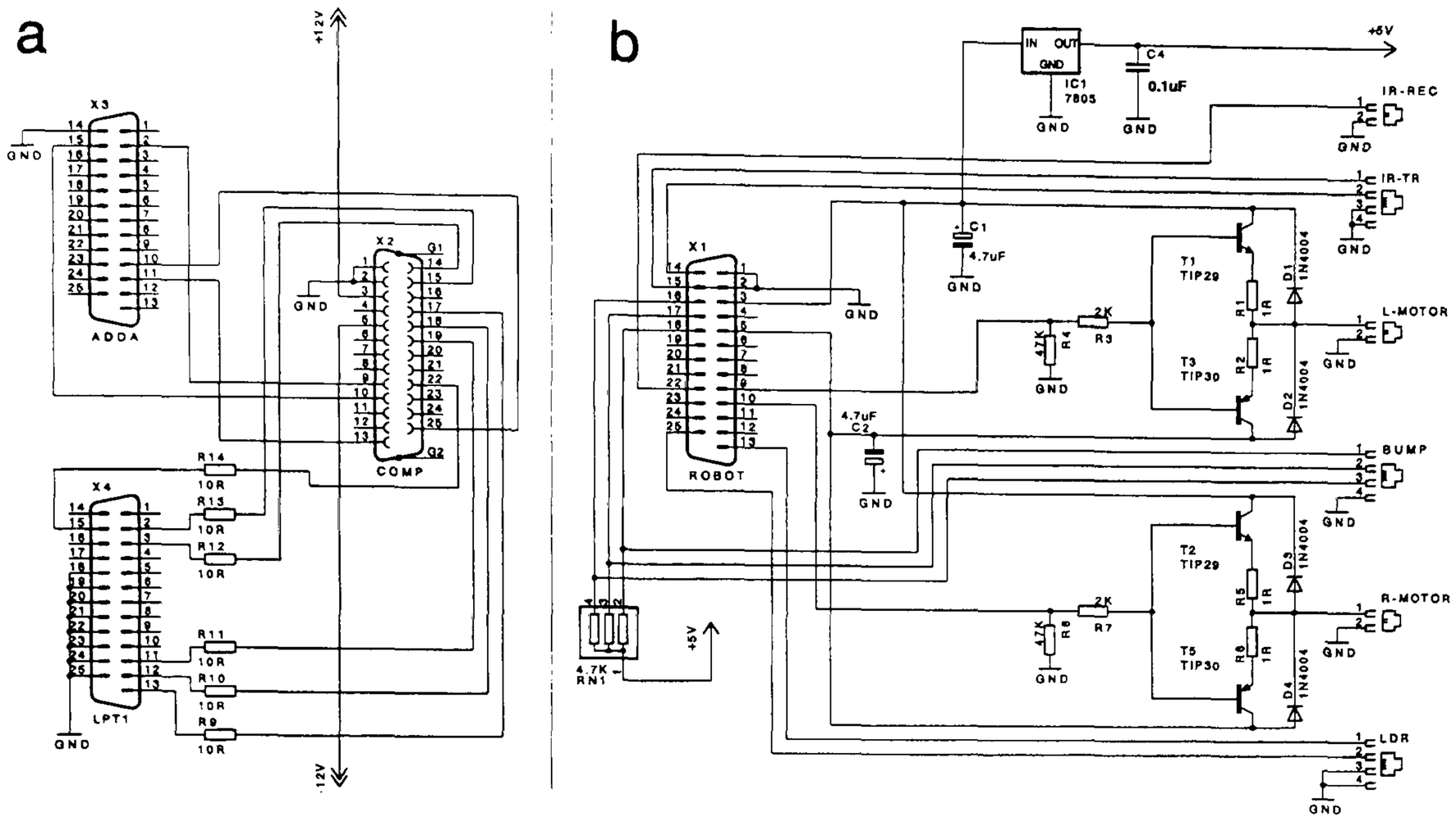

FIGURE B.1: Interfacing between PC (a) and robot (b). Part a) and b) are linked via a cable which is plugged into the connectors X1 and X2. Only those parts are shown which are new compared to the original rug-warrior design. Connector X3 is connected to an AD/DA interfacing card and X4 is connected to the parallel port of the PC. The power ($\pm 12V$) is supplied externally by a standard switching power supply. The bump sensors of the robot (connector BUMP) pull the corresponding line to ground level. For the range-finders the circuitry of the rug-warrior is used. The two infra-red transmitters (IR-TR) are controlled directly by two ports of the printer-port and the signal of the infra-red detector (IR-REC) is directly fed back to the printer port. The D/A converter in the PC provides two analog signals of the range ($-10V \ldots +10V$) which are amplified by T1-T4 and sent to the two motors of the rug warrior. The LDR-signals are for future use.

TABLE B.1: Pinout of the parallel printer-port

Pin	**Name**	**Robot**
2	D0	IR-transmitter, left
3	D1	IR-transmitter, right
11	BUSY	bump, left
12	PE	bump, rear
13	SEL	bump, right
15	/ERROR	IR-receiver

B.1 Motor control

On the PC-side (a) the AD/DA card was used to provide *analog* signals for the motors of the robot. On the robot side two complementary power amplifiers

supply the motors with a maximum current of $250mA$ (limited by R3 and R7). To get even more protection against overcurrents R1,R2,R5 and R6 limit the total current through the transistors to 750mA. Because of the simple design of the power amplifier there exists a dead zone ($\pm 0.7V$) where the input signal causes no output signal (0V). This dead zone was compensated in the control-software. Taking into account the output-range of the DA-converter the active range for the motor was approximately $\pm 8V$.

B.2 Range-finders

The range finders of the robot use a standard IR-receiver which is common in TV-remote controls. Such IR-receivers are only sensitive to pulsed infra-red at a frequency of approx $40kHz$. The 2 IR-transmitters work with such a pulsed frequency and can be switched on and off by the printer-port. To detect obstacles in the 2 directions first one IR-transmitter is switched on and after 1ms the response of the IR-receiver is registered. The same timing protocol applies to the other IR-transmitter which is executed directly after the first one. Thus, the detection takes place within 2ms. Since one time step is 10ms the temporal difference between the left and the right sensor can be neglected. The detection range was adjusted to $0.5 - 15.0\ cm$.

B.3 Bump-sensors

The bump sensors of the robot are directly accessible at the printer-port. Since the bump sensors only pull down the signals to ground an array of three pull-up resistors is used to achieve TTL-level.

Bibliography

Abbott, L. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature Neuroscience supplement*, 3:1178–1179.

Anastasio, T. J. (2001). Input minimization: a model of cerebellar learning without climbing fiber error signals. *NeuroReport*, 12(17):3825–3831.

Anderson, D. (1989). *Artificial Intelligence and intelligent systems: The implications.* Ellis Horwood LTD, Chichester, England.

Arelo, A. and Gerstner, W. (2000). Place cells and spatial navigation based on vision, path integration and reinforement learning. In *Advances in Neural Information Processing Systems*, volume 13, Denver. MIT-Press.

Ashby, W. R. (1956). *An introcduction to cybernetics.* Methnen+Co LTD, London.

Balian, R. (1991). *From microphysics to macrophysics : methods and applications of statistical physics.* Springer Verlag.

Balkenius, C. and Morén, J. (1998). Computational models of classical conditioning: A comparative study. Technical report, Lund University Cognitive Studies 62, Lund. ISSN 1101-8453.

Bellman, R. (1957). *Dynamic Programming.* Princeton University press, Princeton, New Jersey.

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

Bi, G.-q. B. and Poo, M.-m. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.*, 24:139–166.

Blinchikoff, H. J. (1976). *Filtering in the Time and Frequency Domain.* Wiley, New York.

Blum, K. I. and Abbott, L. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Comp.*, 8:85–93.

Bozic, S. M. (1979). *Digital and Kalman filtering: an introduction to discrete-time filtering and optimum linear estimation.* The Gresham Press, Old Woking, Surrey.

Brooks, R. A. (1989a). How to build complete creatures rather than isolated cognitive simulators. In VanLehn, K., editor, *Architectures for Intelligence*, pages 225–239. Erlbaum, Hillsdale, NJ.

Brooks, R. A. (1989b). A robot that walks; emergent behaviors from a carefully evolved network. Technical Report 1091, MIT AI Lab.

Brooks, R. A. (1997). Intelligence without representation. In John, H., editor, *Mind Design II*, chapter 15, pages 395–420. MIT-press, Cambridge, Mass.

Damper, R. I., French, R. L. B., and Scutt, T. W. (2000). ARBIB: An autonomous robot based on inspiration from biology. *Robotics and Autonomous Systems*, 31(4):247–274.

Dayan, P. (2001). Motivated reinforcement learning. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Dayan, P. (2002). Matters temporal. *TRENDS in Cognitive Sciences*, 6(3):105–106.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience.* MIT Press, Cambridge MA.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience supplement*, 3:1218–1223.

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

Der, R. and Liebscher, R. (2002). True autonomy from self-organised adaptivity. In Damper, R. and Cliff, D., editors, *WGW'02. EPSRC/BBSRC International Workshop. Biologically Inspired Robotics — The Legacy of W. Grey Walter*, pages 134–141, HP Bristol Labs, UK. Hewlett Packard.

Descartes, R. (1952). *Descartes' philosphical writings.* Macmilan, London. 1596–1650.

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

Domjan, M. (1998). *Principles of learning and behaviour.* Brooks/Cole, Pacific Grove.

Dorffner, G. (1991). *Konnektionsmus.* Teubner, Stuttgart.

Doya, K. (2000). Reinforcement learning in continous time and space. *Neural Networks*, 12(1):219–245.

Doya, K., Kimura, H., and Kawato, M. (2001). Neural mechanisms of learning and control. *IEEE Control Systems Magazine*, 21(4):42–54.

Ekdahl, B. (2001). How autonomous is an autonomous agent? In *5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information, Systems Analysis, and Synthesis (ISAS 2001)*, volume IX, pages 130–135, Orlando.

Ekström, A., Meltzer, J., and Mc Naughton, B. (2001). NMDA receptor antagonism blocks experience-dependent expansion of hippocampal "place fields". *Neuron*, 31:631–638.

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

Gerstner, W., Kreiter, A. K., Markram, H., and Herz, A. V. (1997). Neural codes: Firing rates and beyond. *Proc Natl. Acad. Sci USA*, 94:12740–12741.

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

Grossberg, S. and Merrill, J. (1996). The hipocampus and cerebellum in adaptively timed learning, recognition and movement. *J. Cogn. Neurosci.*, 8:257–277.

Grossberg, S. and Schmajuk, N. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, 2:79–102.

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

Guo-Quing, B. and Poo, M.-M. (1998). Synaptic modifications in cultured hippocampus neurons. *J Neurosci.*, 18(24):10464–10472.

Haken, H. (1992). *Erfolgsgeheimnisse der Wahnehmung: Synergetik als Schlssel zum Gehirn.* Deutsche Verlags-Anstalt, Stuttgart.

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

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

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

Hawkins, R. D. and Kandel, E. R. (1984). Is there a cell biological alphabet for simple forms of learning? *Psychological Review*, 91(3):375–391.

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

Hilgard, E. R. (1975). *Theories of Learning.* Prentice Hall, Engle-wood Cliffs, New Jersey.

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

Kalman, R. (1960). A new approach to linear filtering and prediction theory. *Trans. ASME, J. Bas. Engineer. Series D*, 82(1):34–45.

Karniel, A. (2000). Human motor control: Learning to control a time varying, nonlinear, many-to-one system. *IEEE Trans. on SMC part C*, 30(1).

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9:718–727.

Kistler, W. M. and van Hemmen, J. L. (2000). Modeling synaptic plasticity in conjunction with the timing of pre- and postsynaptic action potentials. *Neural Comp.*, 12:385–405.

Klopf, A. H. (1986). A drive-reinforcement model of single neuron function. In Denker, J. S., editor, *Neural Networks for Computing: AIP Conference Proceedings*, volume 151 of *AIP conference proceedings*, New York. American Institute of Physics.

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

Koch, C. and Segev, I., editors (1989). *Methods in Neuronal Modeling: From Synapses to Networks.* MIT Press, Massachusetts.

Kosco, B. (1986). Differential hebbian learning. In Denker, J. S., editor, *Neural Networks for computing: AIP conference proceedings*, volume 151 of *AIP conference proceedings*, pages 277–282, New York. American Institute of Physics.

Lettvin, J., Maturana, H. R., McCulloch, W. S., and Pitts, W. R. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47:1940–1951.

Lieberman, D. (1993). *Learning: behaviour and cognition.* Brooks/Cole, Pacific Grove.

Linsker, R. (1988). Self-organisation in a perceptual network. *Computer*, 21(3):105–117.

Lorigo, L., Brooks, R., and Grimson, W. (1997). Visually-guided obstacle avoidance in unstructured environments. In *Proceedings of IROS '97*, pages 373–379, Grenoble, France.

Luhmann, N. (1984). *Soziale Systeme.* Suhrkamp, Frankfurt am Main.

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

Luhmann, N., Maturana, H., Namiki, M., Redder, V., and Varela, F. (1990). *Beobachter.* Wilhelm Fink Verlag.

Mackintosh, N. J. (1974). *The Psychology of Animal Learning.* Academic Press, New York, NY.

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

Maturana, H. and Varela, F. J. (1980). *Autopoiesis and cognition: the realization of the living.* Reidel, Dordrecht.

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

Mele, A. R., editor (1997). *The Philosophy of Action.* Oxford University Press, Oxford.

Meysel, A. (1991). *Mobile Robots.* Wold Scientific, Singapore.

Miller, K. D. (1996a). Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In Donnay, E., van Hemmen, J., and Schulten, K., editors, *Models of Neural Networks III*, pages 55–78. Springer-Verlag.

Miller, K. D. (1996b). Synaptic economics: Competition and cooperation in correlation-based synaptic plasticity. *Neuron*, 17:371–374.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1993). Foraging in an uncertain environment using predictive hebbian learning. *NIPS*, 6:598–605.

Neuhauser, G. L. (1966). *Introduction to Dynamic Programming.* Wiley, New York.

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

Nishiyama, M., Hong, K., Mikoshiba, K., Poo, M.-m., and Kato, K. (2000). Calcium stores regulate the polarity and input specificity of synaptic modification. *Nature*, 408:584–588.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J Math Biol*, 15(3):267–273.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Exp Neurol*, 51:78–109.

Pal, P. and Kar, A. (1996). Mobile robot navigation using a neural network. *Proc IEEE intern. conf. robotics and automat.*, pages 1503–1508.

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

Parsons, T. (1951). *The Social System.* Routledge & Kegan Paul Ltd, London and Henley.

Pavlov, I. (1927). *Conditional Reflexes.* Oxford Univ. Press, London.

Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence.* MIT Press, Cambridge, MA.

Phillips, C. L. (2000). *Feedback control systems.* Prentice-Hall International (UK), London.

Piaget, J. (1930). *The child's conception of physical causility.* Routledge and Kegan Paul.

Quick, T. and Dautenhahn, K. (1999). Making embodiment measurable. In *Proceesings of the 4th Fachtagung der Gesellschaft für Kognitionswissenschaft*, Bielefeld, Germany.

Rachlin, H. (1976). *Behaviour and Learning.* Freeman and Company, San Francisco.

Rao, R. P. and Sejnowski, T. J. (2001). Spike-timing-dependent hebbian plasticity as temporal difference learning. *Neural Comp.*, 13:2221–2237.

Rescorla, R. and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. and Prokasy, W., editors, *Classical Conditioning 2, Current Theory and Research*, pages 64–99. ACC, New York.

Riegler, A. (2002). When is a cognitive system embodied? *Cogn. Syst. Res.*, 3:339–348.

Rieke, F., Warland, D., de Ruyter van Stevenick, R., and Bialek, W. (1997). *Spikes — Exploring the neural code.* The MIT Press, Cambridge, Massachusetts, London, England.

Roberts, P. D. (1999). Temporally asymmetric learning rules: I. Differential Hebbian Learning. *Journal of Computational Neuroscience*, 7(3):235–246.

Scheier, C. and Lambrosios, D. (1996). Categorization in a real-world agent using haptic exploration and active perception. In *From animals to animats*, Cape Cod. Fourth International Conference on Simulation of Adaptive Behaviour, MIT-press.

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

Schultz, W. and Suri, R. E. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comp.*, 13(4):841–862.

Searle, J. R. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences*, 3:417–424.

Segre, A. M. (1988). *Machine learning of Robot Assembly Plans.* Kluwer Academic Publ., Boston.

Sharkey, N. E. and Ziemke, T. (1997). The new wave in robot learnin. *Robotics and Autonomous Systems*, 22(3-4).

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

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

Song, S. and Abbott, L. (2001). Column and map development and cortical remapping through spike-timing dependent plasticity. *Neuron*, 32:339–350.

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919–926.

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

Sutton, R. (1988). Learning to predict by method of temporal differences. *Machine Learning*, 3(1):9–44.

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

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

Sutton, R. S. and Barto, A. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pages 355–378, Seattle, Washington.

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

Traub, R. D. (1999). *Fast Oscillations in Cortical Circuits.* MIT Press, Cambridge.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.

Vaidyanathan, P. (1993). *Multirate Systems and Filter Banks.* Prentice Hall, PTR, Englewood Cliffs, New Jersey.

van Hemmen, J. L. (2001). Theory of synaptic plasticity. In Moss, F. and Gielen, S., editors, *Handbook of Biological Physics*, volume 4, pages 771–823. Elsevier, Amsterdam.

Verschure, P. and Coolen, A. (1991). Adaptive fields: Distributed representations of classically conditioned associations. *Network*, 2:189–206.

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

Verschure, P. F. (1998). Synthetic epistemology: The aquisition, retention, and expression of knowledge in natural and synthetic systems. In *Proceedings of the 1998 IEEE World Congress on Computational Intelligence*, pages 147–153, Anchorage. IEEE.

Verschure, P. F. and Pfeifer, R. (1992). Categorization, representations, and the dynamics of system-environment interaction: a case study in autonomous systems. In Roitblat, H., Meyer, J., and Wilson, S., editors, *Proceedings of the Second International Conference on Simulation of Adaptive behaviour*, pages 210–217, Cambridge. MIT press.

von Foerster, H. (1960). On self-organizing systems and their environments. In Yovits, M. and Cameron, S., editors, *Self-Organizing Systems*, pages 31–50. Pergamon Press, London.

von Foerster, H. (1985). *Sicht und Einsicht: Versuche zu einer operativen Erkenntnistheorie*. Vieweg, Braunschweig.

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.

Walter, W. G. (1953). *The Living Brain*. G. Duckworth, London.

Watkins, C. J. (1989). *Learning from delayed rewards*. PhD thesis, University of Cambridge, England.

Watzlawick, P. (1990). *Menschliche Kommunikation: Formen, Störungen, Paradoxien*. Huber, Bern, 8 edition.

Webb, B. (1995). Using robots to model animals: a cricket test. *Robotics and Autonomous Systems*, 16:117–134.

Widrow, G. and Hoff, M. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, 4:96–104.

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

Wolpert, D. M., Ghahramani, Z., and Flanagan, J. R. (2001). Perspectives and problems in motor learning. *TRENDS in Cognitive Sciences*, 5(11).

Xie, X. and Seung, S. (2000). Spike-based learning rules and stabilization of persistent neural activity. *Advances in Neural Information Processing Systems*, 12:199–208.

Young, D. L. (2001). A hebbian feedback covariance learning paradigm for self-tuning optimal control. *IEEE trans. on SMC. Part B*, 31(2).

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., and Poo, M.-m. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395:37–44.

Ziemke, T. (2001). Are robots embodied? In *First international workshop on epigenetic robotics Modeling Cognitive Development in Robotic Systems*, volume 85, Lund.