Fast Robot Learning Using a Biological Model

Bridget Hallam

<bridget@aifh.ed.ac.uk>

Department of Artificial Intelligence,
The University of Edinburgh,
5 Forrest Hill,
Edinburgh EH1 2QL,
SCOTLAND

Fax: 031-225-9370
Phone: 031-650-1000
or direct dial: 031-650-3080
Telex: 727442 (UNIVED G)

Keywords: learning, motivation, mobile robot, behavioural model, causal learning.

Abstract

An autonomous mobile robot was trained to distinguish between similar stimuli in only four presentations using a biological model of learning originally devised to explain the variety of learning phenomena observed in animals. This paper briefly outlines this model and describes experimental results obtained both from the robot and in simulation. A robot controller based on this model incorporates much designer knowledge of task, agent, and environment but allows most of this information to be adjusted as a result of the robot's experience.

Robot control systems used in the real world tend to be entirely hard-wired, whether through careful planning of the appropriate sequence of actions or through pre-determined reactions to expected stimuli. These systems are normally reliable so long as the environment contains no surprises. Where the environment is less well known in advance, it would obviously be desirable for the robot to learn how to respond to the new stimuli. The problem with robot learning to date is that it generally requires large numbers of presentations of the same stimulus in order to be effective ([Sut84], [MC91] and others). This results in lengthy training times if many different sensor states need to be recognised.

In this paper a biological model of learning is presented which only requires of the order of four presentations of each stimulus situation in order for the appropriate action to be learned. The model was developed to explain the variety of learning phenomena observed in animals and its predictions have been tested on Siamese fighting fish [HD93]. This model responds to temporal sequences of input and generates temporal sequences of actions which affect the environment and therefore the subsequent input pattern. Cause-and-effect sequences are learned in this way. Assumptions about the agent include: many sensor states, multiple connectivity between sensory and behaviour-releasing neurons, and 'enough' behaviours available for the given task.

This model has already undergone preliminary tests as a means for controlling robot behaviour. A simplified version was implemented in an autonomous mobile robot and some experiments performed [HHH93]. These experiments indicated that the model was indeed efficient at associative learning and therefore likely to be useful as a robot controller. The mathematics in the model is 'best guess' fit based on empirical observation. There is scope for a principled examination of the parameter space and of the equations used. This is best done in simulation.

In this paper the version of the model that is currently being used for these tests is outlined. The correlation between results obtained from robot experiments and from the same experiments in simulation is examined. Work done in simulation to investigate the different learning rates which can be achieved and the effect of adjusting the timing of consequence arrival are described and the results discussed. Some of the implications of using the model as a means of controlling robots are then reviewed. Results from these studies will be integrated into the robot control system in due course.

Halperin's Model of Learning and Motivation

Halperin's model [Hal90], [HHH93] comprises a connectionist net composed of three main neuron types arranged in layers. Neurons communicate *via* multiple unidirectional synapses. The first two neuronal layers are multiply connected by plastic synapses; the second and third are recurrently connected in one-to-one correspondence by fixed-weight synapses. Each neuron is more complex than those normally used in neural networks; notably they contain two decaying averaging filters of the form:

$$\frac{dy}{dt} + \gamma\, y \;=\; \gamma \times \text{setpoint} \tag{1}$$

One of these acts to reduce the effect of noise in the input and one to keep a memory of past activity. Synapses are also complex: they continuously keep track of the correlation between the current activity of their post-synaptic neuron and the past activity of their pre-synaptic neuron. This correlation is used to update a weight change function; however weights do not change unless both neurons have been off for a while.

This combination of decaying average processes within the neuron and intermittent synapse weight updating means that the major factor involved in the weight change is the correlation between the *ends* of firing in the neurons — an unusual but valid interpretation of Hebb's rule [Heb49]. It also means that learning cannot occur unless the neurons rest after each bout of activity — an observation which caused Sinclair to formulate his 'rest' principle [Sin81].

The model accounts for many of the behavioural features observed in animals, *e.g.* learning, conditioning of various types, habituation, *etc.*. It also predicted a new type of conditioning called *postponed conditioning*, where an agent learns to respond to a previously neutral stimulus which appears near the end of an inappropriately long behaviour. This postponed conditioning was then investigated and observed in the display behaviour of Siamese fighting fish [Hal90] [HD93]. Therefore the model is biologically plausible in that it explains and correctly predicts actual behaviour in real vertebrate animals.

The model is explained in more detail in [Hal90] and [HHH93].

The three types of neuron are segregated into layers as illustrated in Figure 1. In the simulation, input to the sensory (S) neurons is presented as a set of simple numerical values. On a robot, this input consists of unprocessed sensor readings, logical propositions produced by thresholding sensor data, combinations of such propositions, *etc.*. The input may be raw or processed data, as desired. The amount of input and its temporal sequence is used to determine which S neurons should fire. Output from the S neurons is combined with information encoded in the current synapse weights and passed to the releaser (R) neurons responsible for releasing the behaviours. When a releaser neuron accumulates enough input to fire, its behaviour (B) neuron is stimulated. Competition amongst the B neurons ensures that only one behaviour is active at any time. The winning B neuron provides extra input to its R neuron, creating a positive feedback loop which keeps this behaviour active until the B switches off. B neurons are normally switched off by a new behaviour being triggered; however, they will switch themselves off given time. The behaviour produced is one of a set of short, fixed patterns of actions.

Each S-R synapse has as a parameter the expected time difference t_{exp} between its S neuron finishing firing due to the stimulus disappearing and its R neuron finishing firing due to the successful completion of the behaviour released by that R. Synapse weights increase or decrease according to the correlation between the expected time difference t_{exp} and the observed or actual time difference t_{obs} between S and R finishing firing. The weight update procedure involves a measure of the current average activity of the R neuron and the average activity of the S neuron t_{exp} ago and using these to obtain a value from a 'command surface'. This value is used as the target value for a new decaying average process whose value is then used as the gain for a standard weight change function of the form:

$$\begin{array}{ll} \text{if } \; wt\_ch < 0 & new\_wt = old\_wt \times e^{A \times dt} \\ \text{if } \; wt\_ch \geq 0 & new\_wt = 1 - ((1 - old\_wt)e^{-2A \times dt}) \end{array}$$

where $A = f(s, r)$. Space does not permit a full statement of the weight update differential equations, they are given in [HHH93].

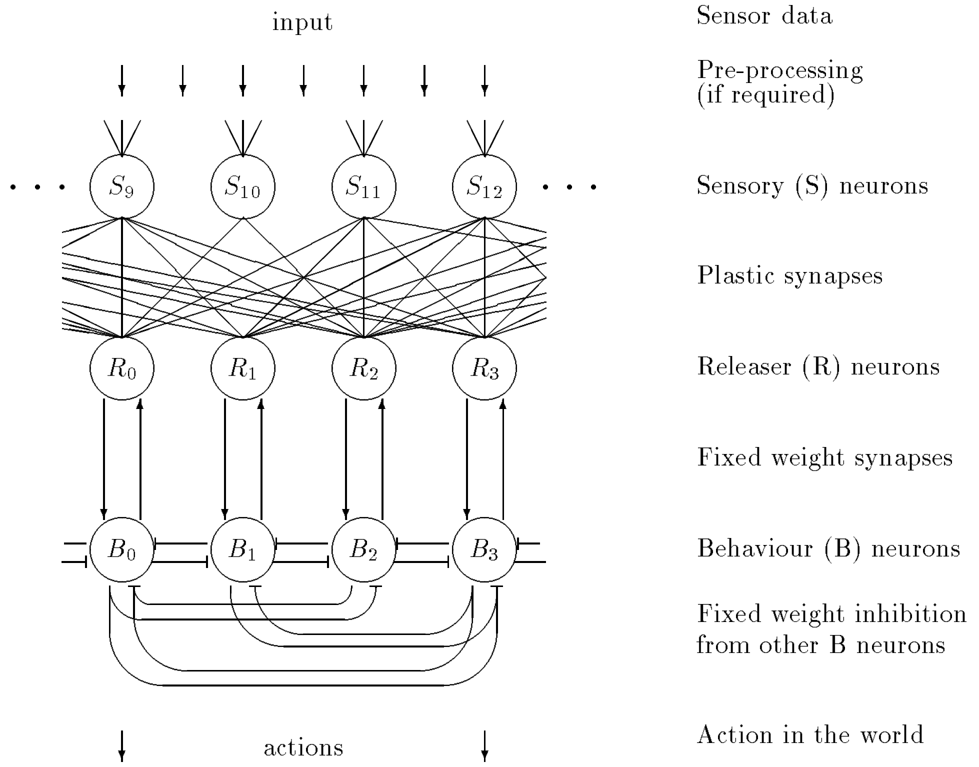

Figure 1: Part of a Halperin Net

The expected time t_{exp} is set such that when a behaviour is successful the correlation between observed and expected differences should be very good and the synapse weight will increase. When a behaviour is unsuccessful the B neuron and therefore its R neuron generally stay on too long, in which case synapses connecting with this R neuron weaken. Weakening is greatest if the observed time is only just outside the expected time, where 'only just' is settable by neuronal parameters. There is little change in synapse weight when the S and R neurons finish firing a long time apart, as the two neuronal firing events are then considered to be completely uncorrelated.

Learning is thus very sensitive to the length of time that any B/R pair stays on after its releasing stimulus disappears. Some B neurons switch themselves and their R off early enough that they always finish firing at about the 'expected time' of some connected synapses. The behaviours they represent are self-reinforcing and act as rewards: ending, confirming and strengthening behavioural sequences. Normally, however, the length of time that a B/R pair stays on is completely dependent upon a new behaviour starting, competition between B neurons causing the current B to switch off. This implies that a new 'releasing' stimulus has been seen (a stimulus that excites an S neuron with a high-weight connection to an R). These new releasing stimuli are generated as a consequence of significant changes in the environment. These changes are normally expected to be caused by the successful action of the current behaviour, but may have other causes. This 'chaining together' of behaviours *via* perception of their consequences means that an agent 'uses the world as its model' in the same sense as, for example, the MIT robot Genghis [Bro90]. Behaviours which are 'not successful' are retried a few times, then are not attempted further. This lack of success backtracks up a behavioural sequence such that eventually the agent will refuse to start a sequence that has rarely proved to be successful.

The implemented version of the model allows fast learning and unlearning of stimulus-response connections as found in some types of classical conditioning and in postponed conditioning. This causes the system to refine its definition of the releasing stimuli for a behaviour to be the most specific and reliable available, which implies that the hand-setting of these releasing connections need only be approximate and the system will determine for itself which stimuli are most reliable.

The model contains numerous parameters and functions which could be adjusted to affect the behaviour of the net, although for most purposes it is appropriate to use one particular set for all neurons. Items which do need hand-setting are: the expected time t_{exp} in each synapse (this will normally be the same for each synapse connected to a given R neuron), the particular sensor conditions under which each S neuron fires, and which S–R connections are to be high weight or releasing connections. The need for hand-setting of releasing stimuli could be removed altogether by comprehensive supervised training of the net. It would also be possible to place a self-organising net between the sensors and the S neurons, so that the agent itself could discover which sensor combinations were of most use to it.

The model described so far contains only the essential elements of the theory of learning and motivation in vertebrates that Halperin proposed. The full version of the model (as described in [Hal90] and [HHH93]) has extra layers of S neurons and considers each S and R neuron described above to consist of a pool of similar neurons with lateral inter-connections. These extra structures comprise a mechanism for the formation of the stimulus-stimulus associations of classical conditioning. The whole net is repeated three times in the full model to form a complex three-dimensional structure with higher level Bs acting by potentiating R neurons in the level below. The top level corresponds to the 'motivational' level in animals, lack of inhibition between the B neurons of this level allowing the simultaneous presence of conflicting motivations.

Halperin is currently working on upgrading the model to use neurons with real-valued output. The experiments described here use neurons with binary output.

Experiments and Results Discussed

The first set of simulation experiments were the same as those performed on the mobile robot [HHH93]. Synapse strengthening and weakening were investigated using the same starting conditions as for the robot experiment. Experiments were run where the observed time difference between the end of firing in the S neuron and the end of firing in the R neuron was varied artificially. From a starting weight of 0.8, the final weight was recorded after one and two presentations of the stimulus. The expected time difference t_{exp} for each synapse was set to be 1.0 second. The results obtained from this investigation were similar to those obtained from the robot, as illustrated in Figure 2.

The results show that the overall shape of the graphs are identical. There is a time window around t_{exp} where the synapse weight has increased during the experiment. Either side of this, where t_{exp} and t_{obs} are just different enough, the synapse weight has weakened. Where t_{exp} and t_{obs} are completely different, *i.e.*where the S and R firing are deemed to be uncorrelated, there is little effect on synapse weight. Note that the maximum level of strengthening achievable is similar to the maximum level of weakening, showing that learning and unlearning can happen at similar speeds. The differences in the two sets of results are likely to be due to inaccuracies in timing for the robot experiment, especially for the short times, and sensor noise. These differences are small enough that conclusions drawn from the results of simulation experiments should prove valid when transferred back on to the robot.

The parameter space of the model was explored in order to investigate what rates of learning could be achieved. With appropriate neuronal time constants, learning could be achieved with as little as one presentation of the stimulus. This is too fast for most applications, as it makes the synapse weight too vulnerable to random changes in the environment. Minor changes to these time constants decrease the learning rate so that more presentations are needed before an association can be said to be formed. These time constants can also be set such that the synapse cannot weaken and the stimulus-response pair cannot be unlearned; such that the synapse cannot strengthen; or such that the strengthening/weakening characteristics are almost the opposite of that illustrated in Figure 2. The maximal unlearning rate can be set lower than the maximal learning rate; we have not been able to demonstrate the reverse while keeping the ability to learn. The ability to have 'unforgettable' stimulus-response connections is a useful feature of the model.

The accuracy required in the correlation between the expected and observed difference in the ends of firing of the synaptic neurons was also investigated. This relates to the width of the time window where strengthening occurs (Figure 2). For any given set of neuronal time constants, the accuracy required is identical whatever

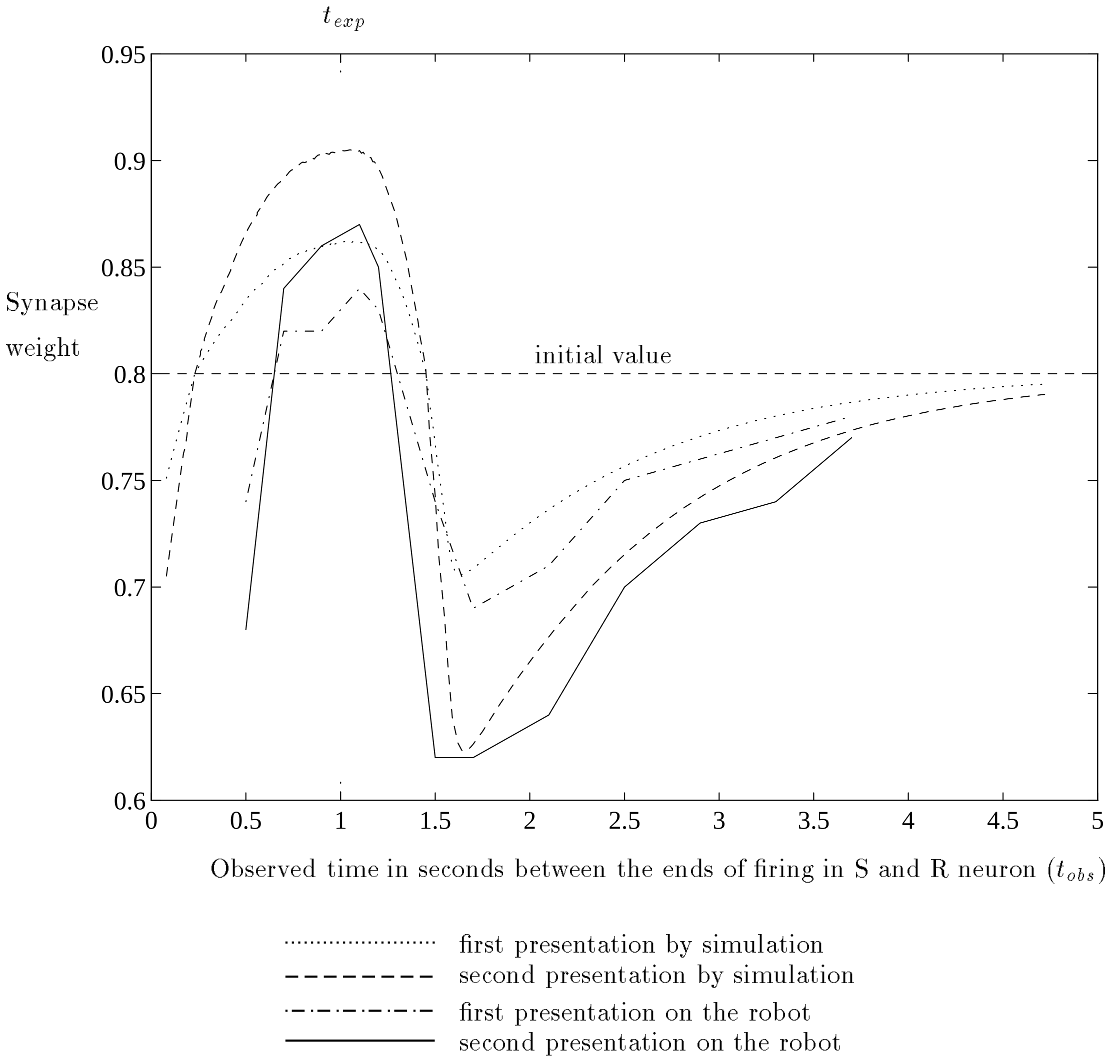

Figure 2: Robot and Simulation Results Compared

the expected time difference. Changing the relationship between certain time constants in the pre- and post-synaptic neurons can affect this function slightly, however most combinations of mismatched time constants do not produce effects consistent with the requirements of the model at all. We have identified the functions in the model which are responsible for the width of the 'strengthening' time window and know the type of changes which are necessary. Substitute equations which allow the width of the strengthening time window to be adjustable for individual synapses without causing the loss of other features of the model are currently being investigated.

The effect on learning of the frequency with which the stimulus is seen was also investigated, by varying the amount of time between presentations and recording the synapse weight after the second presentation. There was found to be no difference in the final synapse weight for interpresentation intervals between 10 and 600 seconds; however intervals of less than 10 seconds did cause the final weight to be lower. For intervals of less than 3 seconds the final weight was less than its initial value. These effects on the final weight are due to there being insufficient time between the two presentations for the weight change to be completed with the particular time constants used.

Postponed conditioning, where a neutral stimulus becomes a releasing stimulus due to its having the correct time relationship with the behaviour, was demonstrated for the robot after three presentations of the training stimulus. Using the same parameters the simulation also showed postponed conditioning after three presentations.

Implications of Using Halperin's Net as a Robot Controller

Halperin's net provides a sensible, general means of connecting perceptions to actions. It requires an agent with many sensor states and behaviours suitable for the task already encoded and working. Agent-specific factors such as the type and number of sensors used and actuators available are dealt with outside the model, as are task-specific factors such as the perception required or the behaviours needed.

The model allows learning to happen extremely quickly under ideal circumstances. Fast learning is particularly advantageous when there might be danger in repeating mistakes, as well as taking less time than traditional reinforcement learning. However, learning which happens too fast can cause learned responses to sensor noise or to unusual combinations of circumstances. One of the great advantages of this model is that the learning rate can be adjusted both globally and locally through neuronal time constants. Some events need this fast learning, others need to be more tolerant of variations in circumstances.

Timing is critical in this model. Behaviours need to finish an invariant time after the 'correct' sensory neuron switches off if the observed time t_{obs} is to be close enough to the expected time t_{exp} for the synapse to strengthen. This can often be achieved through adjusting the propositions in the input pre-processing for the relevant sensory neurons to include not just the releasing stimulus but also something which becomes false during the behaviour. If the behaviour time will be invariant but is not known, then it is possible to use multiple synapses with different t_{exp} between the same two neurons. The ones with times close enough will strengthen and the others weaken. The current system will not support learning if the behaviour time varies by very much, despite this being a stated feature of the model. This problem has been solved in theory and is close to a solution in practise.

Many robot controllers which rely on robot learning ignore the expertise of the designer. On the other hand, most controllers which incorporate detailed designer knowledge require that knowledge to be completely accurate, which cannot be guaranteed for complex situations even discounting the effect of changes in the environment. In contrast, this model incorporates as much designer information as is available but does not require it to be accurate. If there is sufficient computing power for redundant neurons and synapses with different parameters to be used, a 'ball park' figure will suffice. Almost any piece of information can be wrong, so long as the overall effect is to produce the right behaviours the first few times. Those neurons and synapses with the 'right' values will soon become the only ones active. This ability to select 'correct' neurons and 'correct' connections between sensors and actuators is one of the great strengths of this method of robot control. However, the need to update and adjust redundant neurons and synapses increases the cycle time of the program and can lead to responses which are too slow for real-time applications. Pruning of low-weight synapses a short while after the robot had started operation would improve matters here if an initially slow response time would not be dangerous. However, the low-weight synapses are those which allow the system to adjust to change, so should not be pruned too vigorously.

Most real robotic systems currently operational do have a tight limit on the computational power available so cannot run an initial system with many redundant neurons. Therefore the various times and time constants need to be carefully worked out beforehand; by experiment, on paper, or by simulation.

The model in its current form also requires hand-prioritising of behaviours. This is fine for tasks where few behaviours are needed or those required have an obvious priority; however it is impractical for more complex systems. Halperin is currently working on an adjustment to the model which should remove this requirement and allow less constrained action selection.

Finally, this model was produced in order to explain the complexities of animal behaviour and exists in very complex form. Only a very simplified version has been presented here. Much scope remains for producing more complex robot behavioural patterns by using progressively less simplified versions of the model. Scale-up problems are unlikely due to the model being produced in the more complex form first.

Summary

A robot control method which enables fast learning was investigated in simulation. Using the same parameters as were used with the robot, similar results were obtained. Further investigation of the parameter space showed that learning and unlearning rates are easily adjustable both globally and individually for the different 'behaviours' present. For a given set of parameters, the learning rate depends entirely on the number of presentations seen and not on the timing of these presentations once the weight change has been completed.

Learning depends on the correlation between ends of firing in the pre- and post- synaptic neurons. The difference in firing has to be close to the expected time difference (a synaptic parameter) for synapse strengthening to occur. 'Close' should be adjustable according to the control model used, but this will require changes in some equations in the weight update functions. These changes are being investigated.

The control method used was developed as a model of learning in biological systems. This means that the model is more complicated than is needed, which augers well for future generations of robots.

Acknowledgements

Thanks go to the people who reviewed this paper and commented on its intelligibility: my supervisor Gillian Hayes; John Hallam, Bob Fisher, Ashley Lotto, Jeremy Wyatt and Anton Dil.

I acknowledge receipt of SERC grant number 9130847X.

References

[Bro90] Rodney Brooks. Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15, 1990.

[Hal90] Janet RP Halperin. *A Connectionist Neural Network Model of Aggression.* PhD thesis, Department of Ethology, Toronto University, Canada, 1990.

[HD93] Janet RP Halperin and David W Dunham. Postponed conditioning: Testing a hypothesis about synaptic strengthening. *Adaptive Behaviour*, 1(1):39–64, 1993.

[Heb49] Donald Hebb. *The Organisation of Behvaiour.* Wiley, NY, 1949.

[HHH93] Bridget Hallam, Janet RP Halperin, and John CT Hallam. An ethological model of learning and motivation for implementation in mobile robots. Research paper 629, Dept of AI, Edinburgh, 1993. In preparation.

[MC91] Sridhar Mahadevan and Johnathan Connell. Automatic programming of behaviour-based robots using reinforcement learning. In *Proceedings of the 9th National Conference on AI*, volume 2, pages 768–773. MIT press, 1991.

[Sin81] John D Sinclair. *The Rest Principle: A Neurophysiological Theory of Behavior.* Erlbaum, Hillsdale, NJ, 1981.

[Sut84] Richard Sutton. *Temporal Credit Assignment in Reinforcement Learning.* University Microfilms International, 1984.