Monograph:
Marriott, S. and Harrison, R.F. (1995) A Self-Organising State Space Decoder for Reinforcement Learning. Research Report. ACSE Research Report 569 . Department of Automatic Control and Systems Engineering

---

A Self-Organising State Space Decoder for Reinforcement Learning

Shaun Marriott and Robert F. Harrison

Department of Automatic Control and Systems Engineering
The University of Sheffield,
Mappin Street,
Sheffield
S1 4DU, UK.

Research Report No. 569

April 1995

A Self-Organising State Space Decoder for Reinforcement Learning

Shaun Marriott and Robert F. Harrison

Abstract

A self-organising architecture, loosely based upon a particular implementation of adaptive resonance theory (ART) is used here as an alternative to the fixed decoder in the seminal implementation of reinforcement learning of Barto, Sutton and Anderson (BSA). The cart-pole problem is considered and the results are compared to those of the original study. The objective is to illustrate the possibility of controllers that partition state space through experience without the need for a priori knowledge. Input / output pattern pairs, desired state space regions and the network size / topology are not known in advance. Results show that, although learning is not smooth, the reinforcement learning implementation considered here is successful and learns a meaningful control mapping. The adaptive search element and the adaptive critic element of the original (BSA) study are retained.

1. Introduction

Artificial neural networks which learn incrementally by adding new nodes or processing elements during operation have been used to approximate mappings (Platt, 1991; Kandirkamanathan and Niranjan, 1992). This technique obviates many of the problems associated with fixed network structures such as ascertaining the optimum network size/configuration (e.g. Fujita, 1992), deciding upon a connection topology and providing sufficient information capacity (complexity) for adequate representation of the problem domain.

One class of neural network architectures especially suited to increasing learning capacity through experience is that based upon adaptive resonance theory (ART) (Grossberg, 1980; Carpenter and Grossberg, 1987a, 1987b, 1989; Carpenter et al, 1991a, 1991b,1992) ART networks have the capacity of dynamically allocating nodes as required during processing without the need for retraining. This property provides a natural basis for on-line adaptive learning. This paper considers a network based loosely upon some of the principles of ART which acts as a self-organising state space decoder to provide an internal representation of state space.

2. Control systems and the problem of delay.

An area of application for the incremental paradigm is the dynamic partitioning of state space or information space for control and related problems. Very often, it is difficult to establish more than crude qualitative information about state space trajectories on all but the simplest of analytical systems. Ascertaining an accurate model of system dynamics and contriving an objective or cost function (Hocking, 1991) signifying desired behaviour, is usually the preferred route in optimal control problems. Most adaptive methods are indirect and use an estimated system model to recompute controls at each step (Sutton et al, 1992). If adequate knowledge is available, the *a priori* integration of this knowledge into the network can limit the autonomy and flexibility of the network. Autonomous learning systems need to be able to extract and organise information during experience in their particular data rich environment, increasing their information capacity where necessary.

The provision of input-output pairs (consisting of a stimulus and a desired response) by an external teacher is an artificial process which relies upon several underlying presuppositions for its operation as a training method. One such area of presupposition is that of the temporal connection between input-output pairs; this temporal connection forms the basis for state transition dynamics.

The effects of an input on state transitions are not limited to instantaneous changes unless memoryless systems are considered; a more accurate assessment of real world systems is that state transitions are influenced by inputs as a function of the time interval between a particular input and a given state transition. This temporal effect reduces the validity of simplistic stimulus-response pairing of input and output pairs to some extent. Problems which involve delayed feedback to a learning system can be reduced to simple pattern association tasks but require a problem to be solved beforehand by a given teacher in order to specify optimal actions which should be taken by the system (Myers, 1992). One way to take delay into account is to present delayed inputs as part of the pattern pair. However, this requires assumptions about the system model, that is, about how many delayed input terms are required, and thus increases the dimensionality of the input space.

3. The Cart-Pole Problem.

Following Michie and Chambers (1968) and Barto *et al* (1983) the cart-pole system is used to exemplify those aspects of neural networks as autonomous learning systems considered in this paper. The cart-pole system (Figure 1) is a highly non-linear system involving the characterisation of complex state-space trajectories. Classical solution methods require assumptions about the form of the control force function and an objective function (Hocking, 1991). Furthermore, such techniques are rarely general and require an *a priori* analysis of each dynamical system encountered.

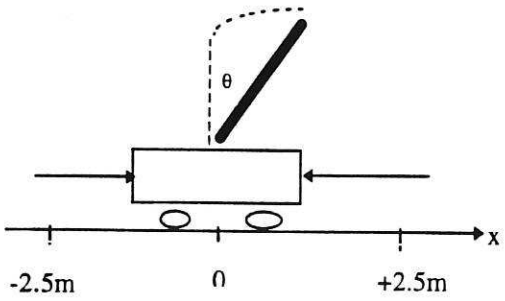

Figure 1. The cart-pole system.

A computer simulation (including friction effects) is used to provide the system. Full details can be found in Barto *et al* (1983). Information from the physical system simulation is minimal and does not provide stimulus-response pairs in the form of inputs and desired control outputs to be associated. Only the state vector and a coarse failure signal, reflecting the cart-pole system status, are supplied to the adaptive control system. If the pole falls or the cart hits the track boundaries then a failure signal is sent to the controller and the cart-pole system is reset to its initial conditions to begin a new trial.

The intention here is not simply to develop another controller for this particular problem; it is to explore some of the issues for which the cart-pole problem provides a convenient example and to indicate the possibilities of developing flexible, general purpose controllers capable of adapting to a given dynamical system through exploring state or information space with a minimum of *a priori* information.

4. Reinforcement Learning

Reinforcement learning (Barto *et al*, 1983; Sutton 1988, 1992; Sutton *et al* 1992; Daynan and Hinton, 1993) arose out of earlier work based upon classical conditioning (Pavlov, 1928; Hebb, 1949; Sutton and Barto, 1981, 1990; Barto and Sutton, 1982; Klopf, 1988). In its simplest formulation it consists of using a single scalar variable representing the punishment / reward status of an artificial neural system with respect to the environment in which it is operating. This signal is fed back to the learning system by a critic which rewards favourable system responses and punishes undesirable ones. Earlier work (Michie and Chambers, 1968) was entirely failure driven. The system considered here is the seminal implementation of Barto, Sutton and Anderson (BSA) (Barto *et al* 1983) which consists of an associative search element (ASE) and an adaptive critic element (ACE); the latter being responsible for interpreting the success / failure status of the ASE subsystem. In the original form, the ASE and ACE are both implemented using a single adaptive element but this is not a necessary condition.

As mentioned in section 2, consequences of actions do not always follow directly from the environment and varying degrees of temporal association have to be taken into account since delays are often present (Myers, 1992). The feedback signal assesses the performance of the network as a whole (Barto *et al* 1983) and thus reflects the importance of the ensemble of actions in eliciting a given environmental (system)

response rather than the single input and instantaneous response implied by learning laws based upon simple associationism.

5. The BSA Implementation

The BSA implementation of a reinforcement learning based controller uses a fixed state space partitioning which gives 162 distinct regions or boxes. A decoder system (Figure 2) assigns a unique output line to each state space region. The set of decoder outputs forms the code which activates a single ASE processing element. During processing, a state vector enters the decoder which switches on the appropriate input line to the ASE which subsequently issues a control action subject to the current system state. Depending upon the outcome and a prediction of reinforcement, the information stored in the activated node, representing the traversed state space region is then updated.

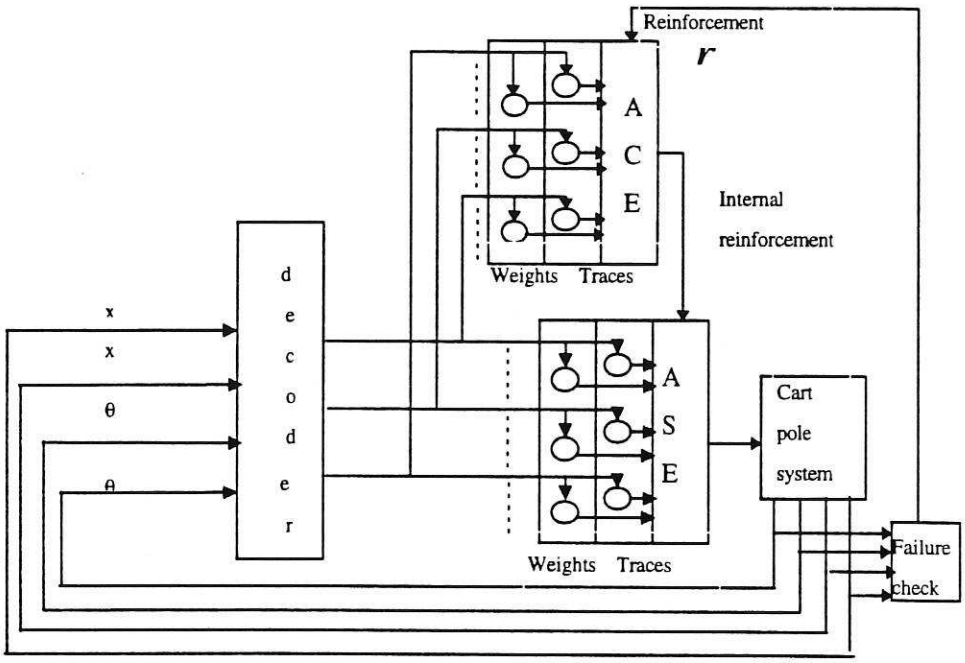

Figure 2. The ASE / ACE reinforcement learning system of Barto *et al* (1983) See text for details.

6. A Self-Organising State Space Decoder

The decoder is a subsystem of the whole control system which lends itself to useful modification. It allows the properties of the controller to be modified whilst retaining the functionality of the ASE/ACE sub-units. Various methods of state space partitioning become possible (e.g. Lin and Kim, 1991) including self-organisation through experience as explored in this paper. Thus, the *a priori* partitioning of state space, as given in the original formulation, is a sufficient but not necessary condition for using the ASE/ACE system.

The decoder system considered here is based upon some of the operating characteristics of one particular implementation of ART, fuzzy ART (Carpenter, Grossberg and Rosen, 1991). To distinguish it from the original architecture, it has

been given the name EUCART. The EUCART decoder uses the Euclidean Metric to establish open balls in a transformed subspace of the state space of $\Re^4$ and uses category bounds based upon the fuzzy ART dynamics to prevent category drift. The latter problem is a common problem of Euclidean based clustering methods (Moore, 1989). EUCART does not include all of the sophisticated mechanisms of fuzzy ART as they are not required for this problem. Like fuzzy ART, the EUCART network input is a vector in the unit hypercube. In this case the cart-pole state vector is scaled by a mapping $\Re^4 \rightarrow [0,1]^4$ and forms the EUCART decoder input. The Euclidean metric, as opposed to a fuzzy metric, was used so that continuity between nearby states was preserved. EUCART partitions state space by forming clusters vectors in the transformed space, $[0,1]^4$ representing state transitions. The decoder begins in a completely naive state with a single node and incrementally increases its representation of state space through experience. Some self-organising systems, (e.g. Kohonen, 1989) require a fixed number of nodes at the outset; this introduces the question of adequate network size as raised in the introduction.

7. The EUCART implementation: Some Results

Simulations comprising 10 runs of 500 trials each, were carried out. The state vector was reset to $x = \dot{x} = \theta = \dot{\theta} = 0$ after each trial. The simulation conditions and parameters were similar to those in the BSA implementation except for a few minor changes necessitated by the new approach. First, runs were not terminated when the trial of a particular run first reached the ceiling of 500,000 time steps of 0.02 seconds (approximately 2.8 hours of simulated time). Learning was still occurring in some cases and the system had to reach the ceiling value a large number of times consecutively to indicate convergence. Second, the learning parameter,α was set to 1,000 in the BSA implementation to establish the control actions quickly. Here, because the state space partitioning was not fixed, learning needed to remain plastic to prevent premature establishment of control actions. Hence α was set to 0.8.

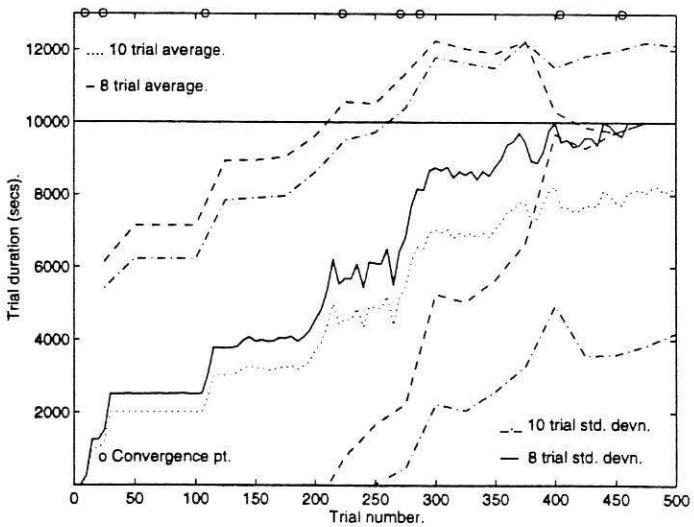

Figure 3. Simulation results showing the performance of the ASE/ACE system with the EUCART state space decoder. The trials were averaged over ten or eight runs. See text.

Figure 3 shows the results of 10 runs and a subset of 8 runs. The subset was required for clarity as 8 of the 10 runs converged to the ceiling value within the 500 trial limit. The solid curve shows the average of the 8 runs which converged during the trial limit. The dotted curve shows the average with the remaining two runs added to the ensembles for each trial. As with the original BSA study, a single point is plotted to indicate the average of each bin of 5 consecutive trial (ensemble) averages. The remaining curves show 1 standard deviation either side of the respective means. These are calculated at 25 trial intervals on the original ensemble values (not on the 5 trial bins). Although the sample size is small, standard deviation is used to indicate spread, since maximum and minimum values are dominated by the trial which converges first. The circles at the top of the graph indicate at which trial the members of the 8 run subset converged.

As expected the trend is towards greater trial durations as the trial number increases. However, the increases in trial duration are not monotonic. This is because the addition of a new EUCART node introduces an initial arbitrary control action. This sometimes pushes the state space trajectory into previously unencountered regions of state space or a region where the control actions are not properly established. The ASE / ACE system is then likely to fail if the well established state space regions are not quickly re-entered. Also, new nodes are sometimes added to cover "gaps" in state space and their influence replaces some well-established state space regions with naive coverage because the regions are now associated with a new node (i.e. the new node centre is now nearer to the regions). As the results show, performance is recovered when the new nodes learn to represent desirable control actions.

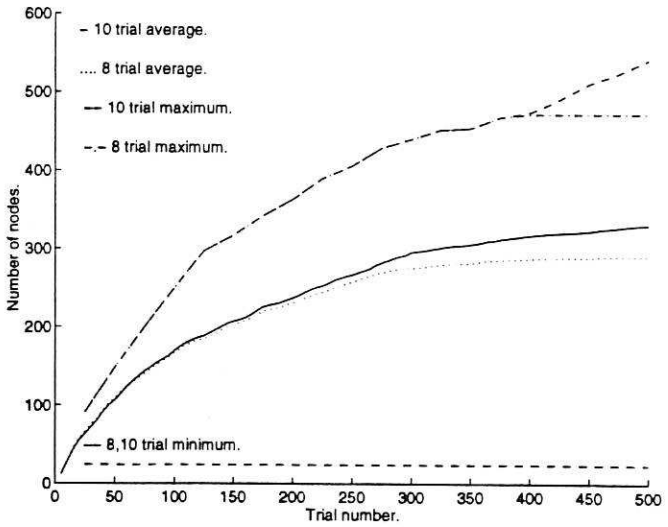

Figure 4. Simulation results showing the increase in the number of EUCART nodes representing individual state space regions and their associated control actions.. See text.

Figure 4 shows the average increase in the number of EUCART nodes for both the full set of 10 runs and for the 8 trial subset. Both the 8 run averages and 8 run maxima reflect convergence to a final set of desirable control actions. The 10 run averages and 10 run maxima indicate that adequate state space coverage has not yet been achieved.

Figure 5 shows one typical run. Again, the results are plotted as an average of bins of five ensemble averages. The graph shows some correlation between increases in node numbers and disruption of trial duration. This is readily apparent at around trial 380 with the small increase in the number of EUCART nodes occurring simultaneously with a drop in the trial duration before recovery and ultimate convergence. The shortest run of the set of trials converges after just 10 trials with only 24 nodes. However, this set of control actions is almost certainly limited as comparatively little of state space has been explored. The controller would not be expected to be as robust and to possess as good disturbance rejection as those controllers with many more nodes indicating a wider experience of state space.

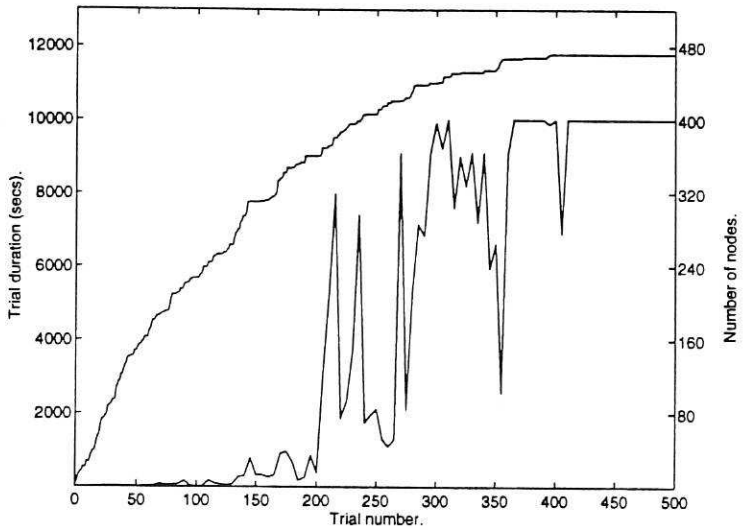

Figure 5. One typical run from the ensemble. Results are plotted as averages of five consecutive trials. Note the transient disruptions caused by the addition of new nodes.

8. Conclusions/ Further Work

The simulations suggest that it is possible to retain Barto, Sutton and Anderson's ASE / ACE subsystems with their proven success while using a self-organising state space decoder. This points to the possibility of autonomous systems which explore state space with few *a priori* conditions for learning. This would make systems more flexible and more generally applicable, although there would be a commensurate increase in initial naivity and thus learning time.

One undesirable feature of the present system is the disruption to average trial duration cased by increases in the number of EUCART nodes. This would possibly be reduced by using information from surrounding nodes to prime the new node thereby obviating the need for arbitrary initial control actions when a new node is created. The experience of the surrounding nodes must be taken into account as nodes with much learning experience are more reliable indicators of successful control actions than naive nodes.

References

Anderson, C. W., (1989) Learning to Control an Inverted Pendulum using Neural Networks, *IEEE Control Systems Magazine,* April

Barto, A. G. (1992) Reinforcement Learning and Adaptive critic Methods, in White, D. A. and Sofge, D. A. (Eds.) Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches.

Barto, A. G. and Sutton, R. S. (1982) Simulation of Anticipatory Responses in Classical Conditioning by a Neuron-Like Adaptive Element, *Behavioral Brain Research*, **4**, pp 221-235.

Barto, A. G. Sutton, R. S., and Anderson, C. W. (1983) Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems *IEEE Trans. Syst. Man. Cybern.* Vol. SMC-13, pp. 834-846.

Carpenter, G. A., & Grossberg, S, (1987a). A Massively Parallel Architecture for a Self-organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing,* **37**, 54-115.

Carpenter, G. A., & Grossberg, S, (1987b). ART 2: Self-organisation of Stable Category Recognition Codes for Analog Input Patterns. *Applied Optics,* **26**, 4919-4930.

Carpenter, G. A., & Grossberg, S, (1989). ART 3: Hierarchical Search Using Chemical Transmitters in Self-Organizing Pattern Recognition Architectures. *Neural Networks,* **3**, 129-152.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B.,(1992). Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Transactions on Neural Networks,* **3**, 698-712.

Carpenter, G. A., Grossberg, S. & Reynolds, J. H., (1991a). ARTMAP: Supervised Real-time Learning and Classification of Nonstationary Data by a Self-organizing Neural Network. *Neural Networks,* **4**, 565-588.

Carpenter, G. A., Grossberg, S., & Rosen, D. B.,(1991b). Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. *Neural Networks,* **4**, 759-771.

Daynan, P. and Hinton, G.E. (1993) Feudal Reinforcement Learning in Henson S.J., Cowan, J. D. and Giles, D. L. (Eds.) Advances in Neural Information Processing, Morgan Kaufmann, San Mateo, CA.

Fujita, O. (1992) Optimization of the Hidden Unit Function in Feedforward Neural Networks *Neural Networks,* **5**, 755-764.

Grossberg, S., (1980). How Does a Brain Build a Cognitive Code? *Psychological Review,* **1**, 1-51.

Hocking, L. M., (1991) Optimal Control: An Introduction to the Theory with Applications, Oxford Applied Mathematics and Computing Science Series, Clarendon Press, Oxford.

Kandirkamanathan, V. and Niranjan, M (1992) Technical Report CUED/F-INFENG/TR.111, Cambridge University, Cambridge, UK.

Klopf, A. H., (1988) A Neuronal Model of Classical Conditioning, Psychobiology, **16**, (2) pp. 85-125.

Kohonen, T.(1989) Self-Organisation and Associative Memory, Third Edition. Springer-Verlag, Berlin.

Lin, C-S. and Kim, H.(1991) CMAC-based Adaptive Critic Self-Learning Control. *IEEE Transactions on Neural Networks.* **2**, 5, pp. 530-533.

Michie, D. and Chambers, R. A.,(1968) BOXES: an Experiment in Adaptive Control, in *Machine Intelligence 2,* E. Dale and Michie, D. Eds. Edinburgh: Oliver and Boyd.

Moore, B., (1989). ART 1 and Pattern Clustering. In Touretzky, D. *et al* (Eds.), *Proceedings of the 1988 Connectionist Models Summer School,* (pp 174-185) San Mateo, CA, Morgan Kaufmann Publishers.

Myers, C. E. (1992) Delay Learning in Artificial neural Networks, Ch. 2, pp 12-28 Chapman and Hall, London.

Pavlov, I. P. (1928) Lectures on Conditioned Reflexes, Vol. I, (Trans.) Gantt, W. H., Lawrence & Wishart Ltd, London.

Platt, J. C. (1991). A Resource Allocating Network for Function Interpolation. *Neural Computation* **3** (2), pp. 215-225.

Sutton, R. S. (1988) Learning to Predict by the Methods of Temporal differences, *Machine Learning,* **3**, pp. 9-44.

Sutton, R. S. (Ed.) (1992) Reinforcement Learning: A Special Issue of *Machine Learning* on Reinforcement Learning Kluwer Academic Publishers.

Sutton, R. S., and Barto, A. G. (1981) Towards a Modern Theory of Adaptive Networks: Expectation and Prediction, *Psychological Review,* **88** (2) pp. 135-170.

Sutton, R. S., and Barto, A. G. (1990) Time-Derivative Models of Pavlovian Reinforcement, in Gabriel, M. and Moore, J. Learning and Computational Neuroscience: Foundations of Adaptive Networks. MIT Press, Cambridge MA.

Sutton, R. S., Barto, A. G. and Williams, R. J. (1992) Reinforcement Learning is Direct Adaptive Optimal control, *IEEE Control Systems Magazine,* April pp 19-22.