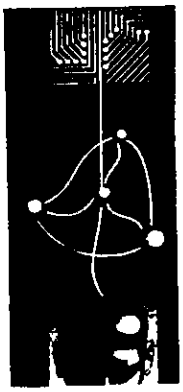


1° Congresso Brasileiro de Redes Neurais

Escola Federal de Engenharia de Itajubá
Itajubá, 24 a 27 de outubro de 1994

I-DYNA: An Architecture for Integrating Reacting, Planning, Learning and Self-Awareness in Autonomous Agents

Camelia Florela VOINEA
Universita' degli Studi di Torino, Italia
Dipartimento di Informatica, cso.Svizzera 185, 10149 Torino
email: camelia@di.unito.it

This work represents a connectionist approach to adaptive reasoning. It proposes a new version of the Dyna class of architectures, named Introspective-Dyna, which is aimed at integrating reacting, planning and learning with self-awareness of an agent in making an action decision. The introspective component of I-Dyna represents for the agent a belief subsystem able to help the agent having an attitude which emerges from within its internal representation of the world. An introspective prediction process is aimed at replacing a belief inferential process. Appropriateness (as soundness) and usefulness (as completeness) of taking an action are judged by means of faithfulness and fulfilment factors, as introspective explanative components of current decision making step. A double-sided temporal difference method is underlying each prediction step: the agent must look-ahead one step and choose an action with most-reachable anticipated rewards (real prediction) and must look backward (at least) one step and choose the action with most-acceptable introspectively anticipated results (introspective prediction). Two version of the I-Dyna algorithm, synchronous and asynchronous, are presented and discussed.

Keywords: reinforcement learning, adaptive reasoning, adaptive neural network, autonomous agents.

1. Introduction

The paradigm of learning from reinforcement is that of the interaction between an agent and its environment in which the agent controls the environment by continuously interpreting at the sensorial level the consequences of its own actions. The agent permanently adapts its reaction by selecting appropriate actions in a sequence which depends on a goal, on time limits and on certain bounds and constraints regarding the physical resources of the environment and of the agent itself. The main theoretical basis of this approach is provided by researches in the area of neural modelling of perception and planning/learning capabilities of adaptive autonomous agents.*Dynamic programming* [Howard, 1960], [Bellman, Dreyfus, 1965] provides an useful but poor model of learning. One of the most important limitations of the dynamic programming is that an action should be tried in order to observe the reward. In the typical case, the reward is associated with the completion of a goal, therefore it could only become available at the end of a sequence of actions. In a sense, dynamic programming methods work backwards from the end of a decision task to its beginning, calculating information pertinent to decision making at each stage based on information previously calculated from that stage to the task's end. As a result of this back-to-front processing, it is difficult to see how dynamic programming can be related to learning processes that operate in real-time as an agent interacts with its environment.*Temporal Differences methods* [Sutton, 1988] can accomplish much the same result, through repeated trials, instead of explicit back-to-front computation. TD methods represent a class of incremental learning procedures specialized for prediction problems. Whereas conventional prediction-learning methods are driven by the error between the real output and the desired one, the temporal difference methods are driven by the error between temporally succesive predictions: with them , learning occurs whenever there is a change in prediction over time.The TD class of prediction methods relies on the simple principle that the observable effects of a sequence of events/actions may be interpretated as ways in which the fundamental laws actually governing the relationships between sequences of events and sequences of observable effects manifest themselves. This principle has been widely applied in pattern recognition and classification. Methods of intelligent control are also using these kind of approach in trying to combine active control with on-going exploration of the behavior of the system to be controled. Combined with a simultaneously formation and adaptation of a world model underlying the prediction process, these methods could provide an appropriate framework for studying the prediction as well as explanative capabilities of embedded systems.*Dyna* [Sutton, 1990] represents a class of simple architectures integrating planning, reacting and learning. Intuitively, Dyna is based on the idea that planning is like trial-and-error learning from hypothetical experience. The Dyna theory is based on the theory of Dynamic Programming, to temporal-difference learning and to AI methods for planning and search. For a fixed policy, Dyna-PI is simply a reactive system, but its policy is continually adjusted by an integrated planning and learning process. The policy is viewed as a plan which is completely conditioned by the current input. The planning process, also called relaxation planning, is incremental and consists of shallow searches, each typically of one step ply, ultimately producing the same result as an arbitrarily deep conventional search. The decision making process is based on continually adjusting the evaluation function in such a way that credit is propagated to the appropriate steps within action sequences.The same algorithm is applied both to real experience (resulting in learning) and to hypothetical experience generated by the world model (resulting in relaxation planning). The results in both cases are accumulated in the policy

and the evaluation function. In spite of its elegance and simplicity, there are however, some weaknesses in the Dyna architecture.The fundamental achievement of Dyna resides, undoubtedly, in the existence of the hypothetical experience itself, in the basic idea of relaxation planning: the use of a "minimum" but, in the same time, "effective" plan about how to act in the next step, the accumulation of both planning and acting experience in the on-going policy and evaluation function with evident improvement of acting process itself, the simultaneity of reacting and planning and the use of the same source of information for updating both reaction and planning in an incremental way. The main weakness of Dyna is that the hypothetical experience does not result also in insight or cognitive capabilities; the agent does not know about the world and about himself more than his evaluation function does, that is, an external measure of his accomplishing a given goal. If this external measure changes, the knowledge the agent has about the world becomes again of no significance. The hypothetical experience in Dyna is meant to express the reasoning of the agent and it has a cognitive aspect inasmuch the plan itself expresses the knowledge the agent has about the world in the current situation, and which is evident by means of the evaluation function. Nevertheless, the agent needs to have an elementary condition of intelligence: self-awareness. How can the agent acquire or be provided with this capability actually represents the aim of this approach.

2. I-Dyna Assumptions

The rationale for these assumptions is that the use of hypothetical experience in prediction requires:(1) a model of the world, particularly a model which is actually formed and updated while the agent interacts with the process to be controlled; the problem is that this kind of model has a high degree of uncertainty in representing the world since it is formed while the world itself is being explored;(2) a theoretical separation, between the input/output provided by the real process and what could represent an input/output for the planning task. Existing approaches on combining reacting and planning in reinforcement learning, especially Dyna [Sutton, 1990], planning with time constraints [Kaelbling, 1991], planning for closed-loop execution using partially observable Markovian decision processes [Chrisman, 1992], have investigated the problem of performing planning in an incremental manner simultaneously with interacting with the process to be controlled. Nevertheless, neither of these approaches suggests an epistemic view of learning by reinforcement, they concentrate mainly on the idea of learning as an incremental task typically as a trial-and-error process and on the optimality principle as a substitute for a more complex belief subsystem able to provide the controller with the capability of interpreting the cognitive value of reinforcement and to have an epistemic attitude with respect to it.

The main working assumptions in this approach are the followings:

1- the choice of an action is influenced by the feed-back provided by the controlled process itself, therefore we will speak about a **closed-loop control architecture**;

2- the input of the decision making process and the internal state have separate representations, since the state cannot be completely observed , so that we will speak about **partially observable Markov processes**;

3- the agent embodies a model of the world and is provided with the capability of reflecting upon the choice of his actions and the effects of his acting by means of an **introspective belief subsystem** which is supposed to express his attitudes towards his own reasoning capabilities.

3. Model Formation and Adaptation in I-Dyna

The particular way in which the present approach makes reference to the idea of insight capabilities an agent could have starts from the idea of "subjective rewards " [Watkins, 1989], and assumes that the action decision making in embedded systems is obviously determined both by the external world and by the internal representation the agent has of his world. The idea of primitive learning [Watkins, 1989] is that an agent could learn from trial-and-error on a stimulus-response base by optimizing his choices with respect to the resources of the environment and to his own resources. Nevertheless, if we want this agent to behave intelligently, we have to accept the idea that it builds a model of its world while interacting with it and, moreover, it has introspective capabilities, that is , it is able to reflect upon the workings of its own cognitive functions.

This approach makes reference to the concepts introduced, from one side, by the epistemologic approach of Gardenfors concerning the dynamics of epistemic states, [Gardenfores, 1988] and, from the other side, by the explanative introspective approch introduced by Konolige [Konolige, 1985]. Upon making the decision on which action to take next, the agent is supposed to make a prediction on the basis of the interpretation he is able to give to the information at hand. This interpretation expresses the knowledge he has in that moment about the state in which he find himself w.r.t the environment and the attitude he takes w.r.t the possible results of taking an action in the given state. This attitude could be expressed also in terms of a prediction in which the subjective rewards guiding the evaluation function are the faithfulness and the fulfillment [Konolige, 1985] factors. This kind of prediction will be further denoted in this paper as *introspective prediction.* The notions of *faithfulness* and *fulfillment* are meant to express here the degrees of epistemic entrenchment [Gardenfores, 1988]. Intuitively, we could imagine the human decision making like a double-linked relationship (see figure 1) between what his goal is and what he believes he can do to achieve this goal:

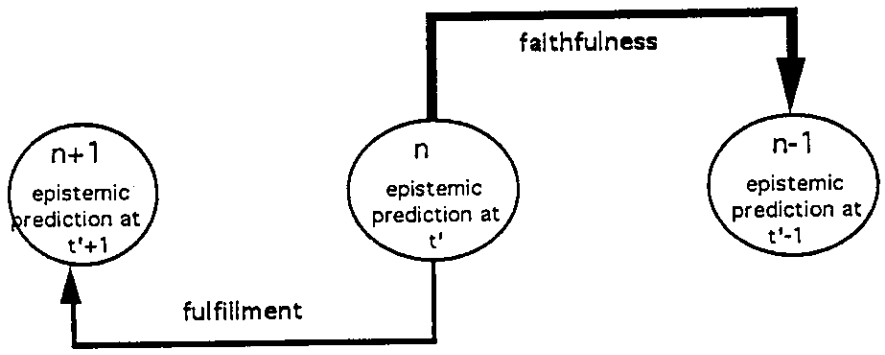


Fig.1. The double-sided temporal difference method: the agent looks-ahead one step to find the most-reachable (really anticipated) action and looks-backward (at least) one step to find the most-acceptable (introspectively posticipated) action.

The *introspective prediction* is a notion denoting the process of recursively generation of the implicit beliefs simultaneously allowed and required by the real prediction process itself. We define this notion for its particular use for understanding the meaning of what precisely I-Dyna architecture embodies as hypothetical experience. The hypothetical experience is a notion used in I-Dyna to denote the introspective beliefs and it is assimilated to some kind of an epistemic commitment function [Gardenfores, 1988]. The input of the introspective module (IP) is represented by a query from P, actually from the real decision making process, addressed to IP.

The output of IP module is an epistemic attitude w.r.t the real prediction performed by the P module (see figure 2). The fundamental relationship between temporarily succesive total predictions (both real and introspective) is given by :

$$E_t \leftarrow E_{t-1} + B^{n+1}$$

where E represents the real prediction at two successive real time moments t and t-1, whereas B represents the introspective prediction at the introspective level (epistemic state) n+1 and at a hypothetical moment of time which will be further denoted by t'. The real prediction E at the current moment t depends on the previous prediction E at t-1 and on what can be derived at a successive introspective level (n+1) if starting from the beliefs at the current introspective level (n) and evaluating the acceptability of what is derived [B at the level n+1] with respect to its base [B at the level n].

The connectionist representation of the introspective prediction allows us to assimilate the degree of epistemic entrenchment [Gardenfors, 1988] to the weights between succesive levels of introspective beliefs (faithfullness and fulfillment factors). The structure of the network itself will provide us with the epistemic commitment function and, moreover, with an explanation of taking one epistemic attitude or another.

The architecture this approach is relying on, Introspective-Dyna, essentially expresses the following idea: the choice of an action in making a decision is the result of combining information provided by the real process itself , I(t), (i.e.: the controlled process, assuming that its outputs are measurable), the agent's knowledge about the process to be controlled, E(t), (expressed in terms of the predictions the agent is able to make upon taking an action in a given state) and the agent's beliefs expressed as introspective predictions, B(n,t'), (i.e: predictions made only on the basis of the previous beliefs about the effects of taking an action in a given situation).

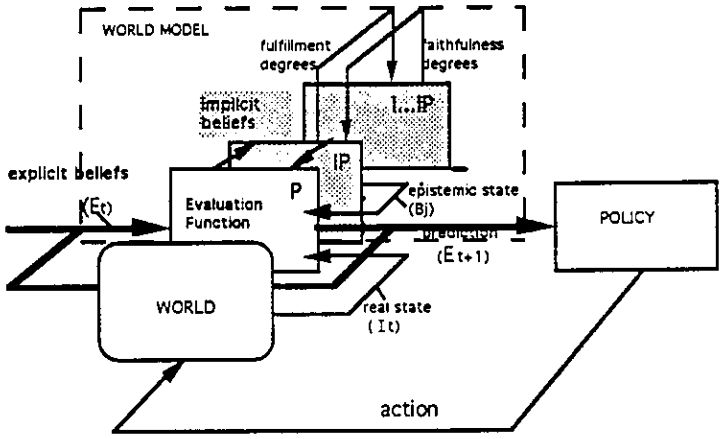


Fig.2. The I-Dyna architecture. Making an acting decision relies on both the world model and the real world. The evaluation function receives both real state information and epistemic state information.

This approach relies, in essence, on the idea that any acting decision an agent could ever make basically expresses his beliefs about his interacting with the world. These beliefs are obviously generated by the model the agent has of its world and by the information (continuously) provided by the world. In making a decision, the agent actually transforms this information into knowledge about the world and about the possible future results (desirable or undesirable) of its interacting with the world.

4. Algorithms

Synchronous I-Dyna Algorithm : *e is the variable for the total prediction and it represents the output, i.e. the action to be taken;ba and breq are the variables for the introspective prediction (the "answer" and the "request", respectively); p is the variable for the real prediction, x,y are real states, r is the reward received from the real process and γ is the discounting factor; n is the variable for the epistemic state, a is the variable for faithfulness factor ;and t is the real time;* initially, e and b coincide; the real time, t, and the introspective time, t', are considered equivalent and simultaneous; each network (RNN - real network, INN - introspective network) is updated by using the double-sided temporal difference method as follows:

for each prediction step: take the action **a** in the real state **x** and introspective state **n**:

update_RNN: $b_{req}(n) \leftarrow i(t) + [p_t(x) + b_a(n-1) + c]$, or

$b_{req}(n) \leftarrow i(t) + e(t)$

$p_t(x) \leftarrow r + \gamma * p_{t-1}(y)$

update_INN: $b_a(n) \leftarrow w_{bb}(n) * b_a(n+1)$

evaluation_function: $e(t+1) \leftarrow e(t) + b_a(n)$

Asynchronous I-Dyna Algorithm: *e is the variable for the total prediction and it represents the output, i.e. the action to be taken;ba and breq are the variables for the introspective prediction (the "answer" and the "request", respectively); p is the variable for the real prediction, x,y are real states, r is the reward received from the real process and γ is the discounting factor; a is the variable for faithfulness factor ;n is the variable for the epistemic state; t is the real time; t' is the introspective time;*

initially, e and b coincide; the real time, t, and the introspective time, t', are not anymore considered equivalent or simultaneous; the relationship between real time and introspective (hypothetical) time is intuitively presented in fig.6 [for each real time prediction (real time unit), the algorithm performs an intrspective prediction over a sequence of n epistemic states, for each state sequence using a corresponding introspective time units; there is, therefore, a 3-level prediction, from which only the first is visible)]; each network (RNN - real network, INN - introspective network, TNN - tendency network) is updated by using the double-sided temporal difference method as follows:

MAKE_REQUEST:

update_RNN: $b_{req}(t,n) \leftarrow i(t) + b_a(n-1,t') + c$

$p_t(x) \leftarrow r + \gamma * p_{t-1}(y)$

MAKE_ANSWER:

update_INN: $b_a(n,t') \leftarrow a(n,t') * b_a(t'+k,n+j)$

GET_ANSWER (PLAN_DELIVERANCE):

if u lower than *Bounds*, then , for some k and some j, such that :

update_TNN: $b_a(n,t') \leftarrow a(n,t') * b_a(t'+k, n+j)$

evaluation_function: $e(t+1) \leftarrow p(t)$ if $t' \ll t$

$e(t+1) \leftarrow p(t) + b_a(n)$ if $t' \equiv t$

and

$n+1 \leftarrow n+j$

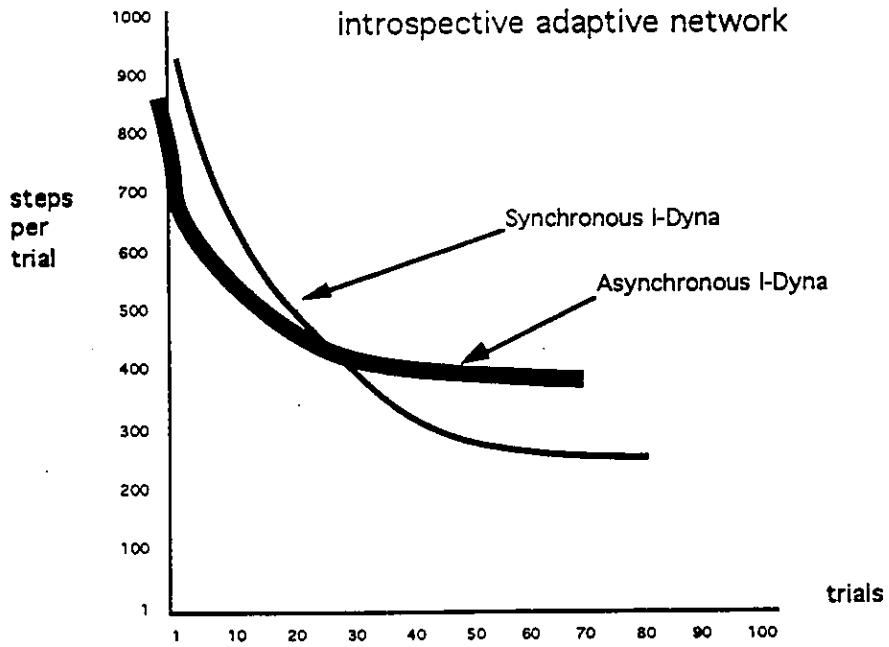
5. Conclusions and Future Work

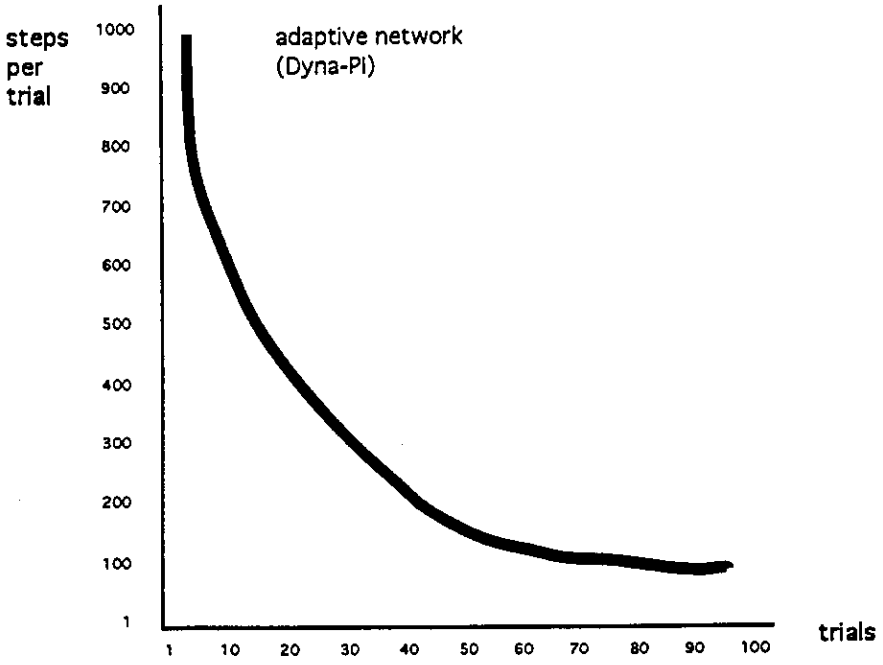
5. Conclusions and Future Work

1. I-Dyna suggests the replacement of an unique external measure of success in decision making with a composed measure of success: internal and external. However, an internal measure of success is bound on the idea of an *internal model of the world*, which comes to the point of "subjective rewards" introduced by Watkins.Trying to overcome the problem of identifying a "content" to these "subjective rewards" scales up to another problem: how does an agent succeed to know its own environment?, what is the precise content of a "cognitive state"?, how does the agent pass from one cognitive state to another ?, and so forth. Thus, the hypothetical experience and relaxation planning in Dyna have inspired another version of Dyna itself: I-Dyna, which is an architecture for integrating reacting, planning and learning with the self-awareness of an embedded system.
2. The idea of relaxation planning has been enhanced in I-Dyna by two types of prediction: prediction of the sensors-measurable future results of taking an action and prediction of the future cognitive attitude (acceptance, refusal) towards the sensors-provided results. It is introduced the notion of *introspective prediction* and its role in substituting a belief inferential process with an introspective prediction process. An *acceptability-function* is therefore recursively constructed from both real data and introspective knowledge.
3. The only prediction rule used here is a temporal difference rule: in order to predict the results of taking an action, the agent must look-ahead (at least) one step and choose the action with most-reachable (anticipated) results, but in order to predict his attitude towards these results, the agent must look-backward (at least) one step and choose the action with most-acceptable (introspectively posticipated) results. We call this *double-sided temporal difference method* .
4. The intention of this kind of approach is to provide the agent with the capability of react and reflect upon its own acting, finding therefore *explanations* for its actions in terms of both the real data and of the model of the environment in which it finds itself.
5. Preliminary results are provided by a *connectionist implementation* of I-Dyna.

References

[**Anderson, 1987**] Anderson, C., W., "Strategy Learning With Multilayer Connectionist Representations", in Proceedings of the 4th International Workshop on Machine Learning, Irvine, California, Morgan and Kaufmann , 1987.
[**Barto, Sutton, Watkins, 1989**] Barto, A., Sutton, R.S., Watkins, C.J.C.H., "Learning and Sequential Decision Making", COINS Technical Report 89-95, September 1989.
[**Barto, Bradtke, Singh, 1991**] Barto, A., Bradtke, S.J., Singh, S.P., "Real Time Learning and Control Using Asynchronous Dynamic Programming", Technical Report 91-57, Amherst, MA: University of Massachusetts, COINS Dept, 1991.
[**Bellman, Dreyfus, 1965**] Bellman, R.E., Dreyfus, S.E., "La programmation dynamique et ses applications", Dunod, Paris, 1965.
[**Chrisman, 1992**], Chrisman, L., "Planning for Closed-Loop Execution using Partially Observable Markovian Decision Processes, from AAAI Spring Symposium Series: Control of Selective Perception, Sytanford, Univ., March, 1992.
[**Gardenfors, 1988**] Gardenfors, P., "Knowledge in Flux. Dynamics of Epistemic States", M.I.T. Press, 1988.
[**Howard, 1960**] Howard, R.A., "Dynamic Programming and Markov Processes", M.I.T. Press, 1960.
[**Kaelbling, 1990**] Kaelbling, L.Pack, "Learning in Embedded Systems", Ph.D. Thesis, Stanford University, U.S.A., 1990.
[**Konolige, 1985**] Konolige, K., "A Computational Theory of Belief Introspection" in Proceedings of the 9th International Joint Conference on Artificial Intelligence, 18-23 August, 1985, Los Angeles, California, A.Joshi (Ed.), Morgan Kaufmann Publishers..
[**Millan, Torras, 1992**] Millan, J. del R., Torras, C., "A Reinforcement Connectionist Approach to Robot Path Finding in Non-Maze-Like Environments", in Machine Learning Vol. 8, No. 3-4, May 1992, Kluwer Academic Publishers.
[**Rabiner, 1989**] Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol.77, No.2, February 1989.
[**Singh, 1992**] Singh, S.P., "Solving Multiple Sequential Tasks Using a Hierarchy of Variable Temporal Resolution Models", Machine Learning Conference, 1992.
[**Sutton, Barto, 1981**] Sutton, R.S., Barto, A., "An Adaptive Network That Constructs and Uses an Internal Model of Its World", in Cognition and Brain Theory, Vol. IV, No.3, 1981, Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.
[**Sutton, Pinette, 1985**] Sutton, R.S., Pinette, B., "The Learning of World Models By Connectionist Networks", in Proceedings of the 7th Annual Conference of the Cognitive Sciences Soc., 1985.
[**Sutton, 1988**] Sutton, R.S., "Learning to Predict by the Methods of Temporal Differences", in Machine Learning No.3: 9-44, Kluwer Academic Publishers, Boston, 1988.
[**Sutton, 1990**] Sutton, R.S., "Integrated Architectures for Learning, Planning and Reacting Based on Approximating Dynamic Programming", in Proceedings of the 7th International Conference on Machine Learning, June, 1990.
[**Sutton, 1991**] Sutton, R.S., "Planning by Incremental Dynamic Programming", in Proceedings of the 9th International Workshop on Machine Learning, 1991, Morgan and Kaufmann.
[**Voinea, 1994**], Voinea, C.F., "Model Formation and adaptation in Reinforcament Learning", in Proceedings of the International Conference on Neural Modeling, University of Lyon, Lyon, France (to appear).
[**Watkins, 1989**] Watkins, C.J.C.H., "Learning from Delayed Rewards", Ph.D. Thesis, King's College, 1989.





Learning Performances of the Introspective Adaptive Network.