Adaptation by Reinforcement Learning in Cooperative Autonomous Robots

Maja J Matarić
Volen Center for Complex Systems
Computer Science Department
Brandeis University
Waltham, MA 02254
tel: (617) 736-2708
fax: (617) 736-2741
email: maja@cs.brandeis.edu

Abstract

We demonstrate the complexities in designing and evaluating Alife learning systems using physical robots interacting in complex, dynamic environment. We use the examples of two different implemented multi–robot learning systems and their learning data on different tasks to illustrate the issues. We describe the difficulties with traditional "objective" methods of evaluation which fail to take into account the complexities of the system and illustrate them on each of the robot learning experiments. We then summarize the evaluation methods we used to evaluate and describe the system behavior without an *a priori* experimenter bias.

Keywords: adaptive robots, reinforcement learning, autonomous agents, swarm intelligence, multi–agent systems, division of tasks.

1 Introduction

Adaptation is one of the hallmarks of biological systems, and one of the main challenges of Alife, a field striving to produce artificial systems which display properties of biological life. Adaptation is particularly difficult yet crucial in systems dealing with dynamic, uncertain environments, such as the physical world mobile robots interact with. The field of robotics has worked on achieving adaptive behavior through various approaches, from traditional adaptive control systems (Barto 1990), to more recent neural–networks (Pomerleau 1992, Millán 1994) and genetic algorithms (Floreano 1993, Steels 1994). Many of those approaches fall under the broad and, in the last decade in particular, exceedingly popular approach called "reinforcement learning."

Learning by reinforcement is widely spread in biology, and animals use it to effectively fine–tune old behaviors and acquire various new ones (McFarland 1985). However, studies in ethology and psychology show that learning curves of animals and artificial agents differ drastically, not only in speed to convergence, but their basic nature (McFarland 1987, Matarić 1994*a*). In this paper we discuss some reasons underlying these fundamental differences.

2 Choice of Domain and Related Work

According to evolutionary theory, the complexity of the niche drives the adaptation of the species. Many fundamental problems with adaptation in artificial systems do not appear until the systems and their environment are sufficiently complex. Even the most benign physical environments offer abundant potential for complexity due to the dynamics interaction between the simplest of physical robots with the environment (Smithers 1994).

Another challenging and critical level of complexity is introduced when multiple agents interact in the physical world, and nature abounds with examples of such systems. Much research has addressed societies with large numbers of simple agents, usually referred to as "swarm intelligence", dealing with the design of communication protocols, social rules, and strategies for avoiding conflict and deadlock (Fukuda, Sekiyama, Ueyama & Arai 1993, Dario & Rucci 1993, Dudek, Jenkin, Milios & Wilkes 1993, Huang & Beni 1993), etc.

Many Alife researchers have focused on modeling colonies of ant–like agents (Corbara, Drogoul, Fresneau & Lalande 1993, Colorni, Dorigo & Maniezzo

1992, Drogous, Ferber, Corbara & Fresneau 1992, Travers 1988), and some have experimented with both live and simulated ants (Deneubourg & Goss 1989, Deneubourg, Goss, Franks, Sendova-Franks, Detrain & Chretien 1990, Deneubourg, Theraulax & Beckers 1992). Related work, including Assad & Packard (1992) and Hogeweg & Hesper (1985), demonstrated simulations of other simple organisms producing complex behaviors emerging from simple interactions, and the evolution of simple communication strategies in such systems (Werner & Dyer 1990, MacLennan 1990).

Until recently, work in Alife largely focused on systems with large numbers of extremely simple agents. In the last few years, several systems featuring multiple mobile robots have been used to implement and test Alife methodologies (Matarić 1992, Beckers, Holland & Deneubourg 1994). Experiments using such systems have brought to light many important issues and have brought into question a number of basic assumptions held by most of the research community. This paper focuses on those issues concerning adaptation with reinforcement learning, currently the most popular on–line learning method for autonomous agents.

3 Reinforcement Learning

Reinforcement learning (RL) is a class of learning methodologies in which the agent adapts based on external feedback it receives from the environment, and interprets as positive or negative scalar reinforcement. The agent's goal is to maximize reward and minimize punishment over time, by producing an effective mapping of states to actions called a *policy.* Traditionally, the agent is given no explicit built–in knowledge about the task. If it receives no direct instruction or answers from the environment the learning is considered *unsupervised* (Barto 1990).

Reinforcement learning has been applied to a variety of domains and problems, including maze–learning (Minsky 1954), checkers (Samuel 1959), the Bucket Brigade algorithm in Classifier Systems (Holland 1985), and Temporal Differencing (Sutton 1988). It has been implemented with with a variety of algorithms ranging from table–lookup to neural networks. Most recently, RL has been attempted on situated, embodied agents. The next section briefly outlines and summarizes the assumptions that were challenged in this new domain. The remainder of the paper uses robot experimental data to illustrate the enumerated issues.

4 Challenged Assumptions

The following are basic assumptions integral to theoretical reinforcement learning which are challenged in complex, embodied adaptive robot domains:

- **Markovian Learning Models:** Most learning algorithms are based on Markovian models that assume the world and the agent can be represented as synchronized finite state automata, with the agent choosing actions and the world neatly changing states in response. In physical domains none of those assumptions hold true due to incomplete knowledge, limited and faulty perception, noise, and intrinsic uncertainty (Matarić 1994*b*).

- **Probabilistic Learning Models:** Alternatives to Markovian models have been proposed. They feature probabilistic state transitions, but require world models that specify the probabilities, and which are unavailable for physical robot domains that are extremely difficult to characterize precisely. As was argued in Matarić (1994*a*), producing such models requires observing the system long enough to establish sufficient statistics, thus long enough to learn an adaptive behavior.

- **Asymptotically Converging Algorithms:** Most successful traditional reinforcement learning algorithms are based on dynamic programming (Bellman 1957) and are favored because of their provable asymptotic convergence properties (Barto, Bradtke & Singh 1993). However, these algorithms traditionally trade built–in knowledge for large numbers of learning trials. The opposite appears to be true for animals, whose reinforcement learning curves are much steeper and converge with many fewers iterations (McFarland 1985). This is not surprising, as evolution favors faster adaptation to dynamically changing environments. Additionally, evolution employs "parallel search" for effective solutions through large populations where individuals are disposable. In contrast, for pragmatic reasons, physical robotic agents must conserve individuals. Furthermore, robot learning must minimize the number of required trials, even in learning behaviors that are, to date, relatively simple compared to any biological ones.

- **Optimal Policies:** Theoretical models allow for establishing a canonical optimal policy or behavior which can be used to evaluate the agent's performance. In physical environments, however, it is typically impossible to specify a single, objective optimal policy. While some behaviors are clearly better than others, many appear "good enough" or similar enough in performance relative to the goal.

Adaptive autonomous agents do not have a single well–defined *goal* in a traditional sense. Instead, like animals, they maintain large numbers of maintenance goals, including avoiding damage to themselves and staying charged. In addition, they pursue multiple achievement goals which change over time, such as finding food, shelter, and information. Achievement and maintenance goals are difficult to translate into a single monolithic evaluation function. In order to combine them, high–level goals must be represented in terms of low–level actions, a process that greatly enlarges the agent's search space, and thus requires even more learning trials (Matarić 1994*a*). In worlds with continuous space, time, and goals, a single optimal strategy is difficult to define and translate all the way down to the action level.

• **Objective Evaluation Strategies:** Related to the challenges of defining an optimal policy, it is generally very difficult to find ways to objectively evaluate the performance of a physically embodied adaptive system. While it may appear quite simple to define evaluation criteria in terms of "time to convergence" and "energy spent", it is usually not helpful to measure absolute time in event–driven systems that can vary greatly from one–experiment to another, as is the case in multi–agent systems. Furthermore, simple measures of energy, such as battery power, can produce irrelevant results, since they hide important details about the environment and system dynamics that cannot be represented in the evaluation function. This and the other issues we have addressed will be illustrated on robot data in the following sections.

5 Experimental Environments

We have employed two different multi–robot experimental environments. The first consisted of four mobile robots capable of detecting and moving objects, and sensing and communicating with other robots (Figure 1). The second environment involved a pair of communicating six–legged robots capable of sensing an object in front of them through contact (Figure 2). In both environments, the robots were equipped with on–board power and computation. Their sensory capabilities provided only local, partial information about the world, the surrounding objects, and other robots, and allowed only for

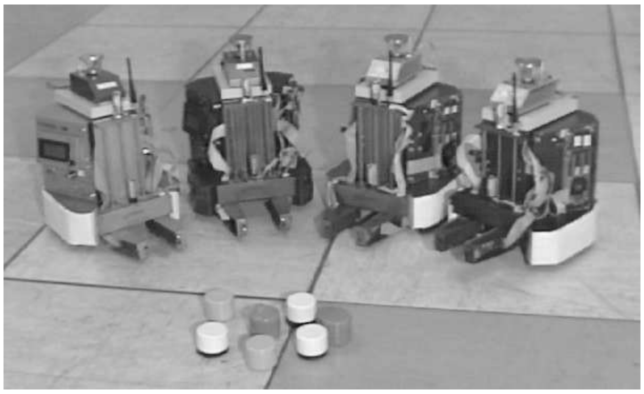

Figure 1: The mobile robots used to validate the group behavior and learning methodologies. These robots demonstrated learning to forage by using group safe–wandering, following, and resting behaviors.

undirected local radio communication. The systems were programmed in the Behavior Language, a parallel distributed robot programming language based on the Subsumption Architecture (Brooks 1989, Brooks 1990).

Our work was motivated by the challenge of learning through social interaction in which agents learn in parallel and independently, while the society as a whole dynamically adapts.

6 Learning in Cooperating Wheeled Robots

The following is a brief summary of the learning system we implemented. A detailed description of the experiments and the data can be found in Matarić (1994*a*) and Matarić (1994*b*).

The wheeled robots were used to implemente a learning behavior consisting of learning to forage. We proposed two forms of shaped reinforcement, heterogeneous reward functions and progress estimators, both of which accelerated learning. We used a simple learning algorithm that summed the reinforcement over time and no discounting, since the policy to be learned was purely reactive (Matarić 1994*b*).

The behavior space included the foraging subset of basic behaviors, augmented with *grasping*, *dropping*, and *resting*. The condition set was reduced

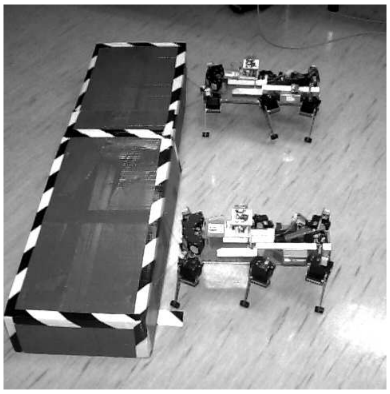

Figure 2: Two Genghis-II six–legged robots were used in the experiments. The robots are 30 cm long and 15 cm high. The sensory suite of each robot consists of two frontal contact whiskers and five pyroelectric sensors for detecting moving IR sources. The robots are powered by rechargeable batteries and controlled by five 68HC11 microprocessors.

to the power set of: *have-puck?*, *at-home?*, *night-time?*, and *near-intruder?*. Three learning algorithms were tested; the first was standard RL Q-learning, the second used heterogeneous reward functions (HRFs), and the third used HRFs and progress estimators. The performance of each of the three algorithms was averaged over 20 trials and plotted based on the mean percent of correct policy learned (Figure 3).

As can be seen from the histogram, our proposed algorithm outperforms the alternatives, but that only proves our hypothesis without giving us any insight into the reasons. The traditional method for data analysis, the comparison to the optimal policy, was of little use here, even though we had an optimal policy available from our previous experiments. In order to understand the results, we tabulated the data for each condition–behavior (state–action) pair and evaluated it based on three additional criteria: 1) time to convergence to a any solution, 2) stability of the solution, and 3) correctness of the solution. We found that some learned conditions oscillated between two similar solutions one of which we had considered to be "correct" but both of which turned out to be equally effective for the agent. Only after an extensive qualitative analysis of each entry in the learning table were we able to produce an evaluation of

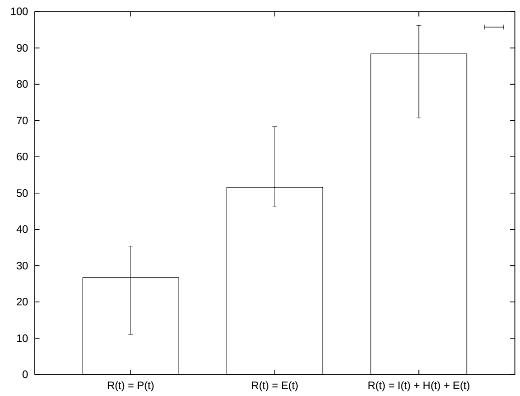

Figure 3: The performance of the three reinforcement strategies on learning to forage. The x-axis shows the three reinforcement strategies. The y-axis maps the percent of the correct policy the agents learned, averaged over twenty trials.

the results, as described in Matarić (1994*b*).

7 Learning in Cooperating Legged Robots

In the case of the legged robots, the learning problem consisted of finding the best strategy for cooperatively moving an oversized elongated box to a goal indicated with a light. The task was challenging because it involved complex real–world dynamics of the robot–to–robot and robot–to–box interactions. The size of the box prohibited efficient individual solutions and required that the robots a coordinated cooperative strategy (Matarić, Nilsson & Simsarian 1994). The input space consisted of ten pyroelectric sensor inputs, five from each robot.

Two different learning tasks were implemented. In one, the the robots were given a hard–wired policy for what to do (*my–action*) in each state, selecting from the set: *left*, *right*, *forward*, and *stop*, but had to learn what to communicate (*its–action*) to to the other robot in order to coordinate their actions. In the other task, the robots had to learn both what to do and what to communicate in each situation. In both cases, the robots used their whiskers to align to the box, and the learning was based on the inputs from

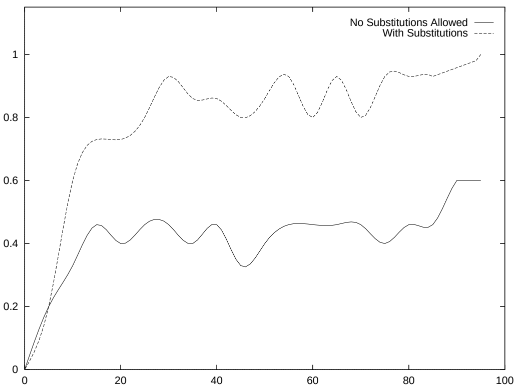

Figure 4: This graph illustrates the performance of concurrent learning of both robot–action matrices, *my–action* and *its–action*. The x–axis plots contiguous trials over a 50–minute run. The y–axis plots the percentage of the learned policy. The solid graph shows the percentage of correct values relative to the optimal policy. The dashed graph shows the percentage compared with a relaxed margin of error.

the pyroelectric sensors indicating the robot and the box position relative to the light at the goal. We also tested different communication strategies, one using only individual sensory information, and the other using the information communicated from the other robot Mataric & Simsarian (1995).

Even more so than in the puck–collecting example, the box pushing task made it difficult to establish a monolithic and unique optimal policy because many combinations of cooperative actions and trajectories would result in efficient box pushing. We evaluated the results using two different policies. In the first case we compared the learned matrix to the actions that appeared to most quickly take the object to the goal. In the second case we relaxed the policy matrix to include substitute actions discovered by the robots through their interaction with the environment during the learning task.

As can be seen in Figure 4, showing the comparison between the strict and relaxed evaluation policy, the difference in the system "performance" is significant. If we merely observe the "objective" performance, we get a different evaluation than if we use the one discovered by the robots for the same goal. In both cases the box is delivered and the goal is reached, so it is difficult to select one as preferable and more objective. This brings up an old trade–off between the amount of built–in bias the designer provides and the time it takes the system to converge. The less is built–in the less biased the system is but the slower it learns. We opted for a simple shaped reinforcement approach that allowed the system to find its own alternative strategy that effectively reached the goal (Mataric & Simsarian 1995). In turn, we had to use an appropriate evaluation methodology in order to properly assess the system's learning performance.

8 Conclusions

In this paper we have discussed some of the key issues in designing and evaluating adaptive Alife systems using physical robots interacting in complex, dynamic environments. We used the examples of two different multi–robot learning systems to illustrate the difficulties with traditional "objective" evaluation methods. Finally, we summarized the evaluation methods we used to properly describe the system behavior. In summary, as the complexity of the environment and the agent interaction increases, so will the challenges of the learning algorithms even for the simplest of tasks. Consequently, it will continue to be necessary to look beyond simple numerical recipes toward more detailed and qualitative evaluation in order to describe all of the system behavior.

Acknowledgements

The author wishes to thank Kristian Simsarian for collaborating on the two–robot box–pushing project, for patiently running experiments, and for analyzing the pushing data. Many thanks to Martin Nilsson for writing and debugging the communication hardware. Thanks to both for many inspiring discussions about all topics from robot hardware to learning.

References

Assad, A. & Packard, N. (1992), Emergent Colonization in an Artificial Ecology, *in* F. Varela & P. Bourgine, eds, 'Toward A Practice of Autonomous

Systems: Proceedings of the First European Conference on Artificial Life', The MIT Press, pp. 143–152.

Barto, A. G. (1990), Some Learning Tasks from a Control Perspective, Technical Report COINS TR 90–122, University of Massachusetts.

Barto, A. G., Bradtke, S. J. & Singh, S. P. (1993), 'Learning to Act using Real-Time Dynamic Programming', *AI Journal*.

Beckers, R., Holland, O. E. & Deneubourg, J. L. (1994), From Local Actions to Global Tasks: Stigmergy and Collective Robotics, *in* R. Brooks & P. Maes, eds, 'Artificial Life IV, Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems', The MIT Press.

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

Brooks, R. A. (1989), 'A Robot that Walks; Emergent Behavior from a Carefully Evolved Network', *Neural Computation*.

Brooks, R. A. (1990), The Behavior Language; User's Guide, Technical Report AIM-1127, MIT Artificial Intelligence Lab.

Colorni, A., Dorigo, M. & Maniezzo, V. (1992), Distributed Optimization by Ant Colonies, *in* F. Varela & P. Bourgine, eds, 'Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life', The MIT Press, pp. 134–142.

Corbara, B., Drogoul, A., Fresneau, D. & Lalande, S. (1993), Simulating the Sociogenesis Process in Ant Colonies with MANTA, *in* 'Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life', pp. 224–235.

Dario, P. & Rucci, M. (1993), An Approach to Disassembly Problem in Robotics, *in* 'IEEE/TSJ International Conference on Intelligent Robots and Systems', Yokohama, Japan, pp. 460–468.

Deneubourg, J.-L. & Goss, S. (1989), Collective Patterns and Decision-Making, *in* 'Ethology, Ecology and Evolution 1', pp. 295–311.

Deneubourg, J. L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C. & Chretien, L. (1990), The Dynamics of Collective Sorting, *in* 'From Animals to Animats: International Conference on Simulation of Adaptive Behavior', The MIT Press, pp. 356–363.

Deneubourg, J. L., Theraulax, G. & Beckers, R. (1992), Swarm-Made Architectures, *in* F. Varela & P. Bourgine, eds, 'Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life', The MIT Press, pp. 123–133.

Drogous, A., Ferber, J., Corbara, B. & Fresneau, D. (1992), A Behavioral Simulation Model for the Study of Emergent Social Structures, *in* F. Varela & P. Bourgine, eds, 'Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life', The MIT Press, pp. 161–170.

Dudek, G., Jenkin, M., Milios, E. & Wilkes, D. (1993), A Taxonomy for Swarm Robotics, *in* 'IEEE/TSJ International Conference on Intelligent Robots and Systems', Yokohama, Japan, pp. 441–447.

Floreano, D. (1993), Patterns of Interactions in Shared Environments, *in* 'Toward A Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life', pp. 347–366.

Fukuda, T., Sekiyama, K., Ueyama, T. & Arai, F. (1993), Efficient Communication Method in the Cellular Robotics System, *in* 'IEEE/TSJ International Conference on Intelligent Robots and Systems', Yokohama, Japan, pp. 1091–1096.

Hogeweg, P. & Hesper, B. (1985), 'Socioinformatic Processes: MIRROR Modelling Methodology', *Journal of Theoretical Biology* **113**, 311–330.

Holland, J. H. (1985), Properties of the bucket brigade algorithm, *in* 'Proceedings, International Conference on genetic Algorithms and Their Applications', Pittsburgh, PA, pp. 1–7.

Huang, Q. & Beni, G. (1993), Stationary Waves in 2–Dimensional Cyclic Swarms, *in* 'IEEE/TSJ International Conference on Intelligent Robots and Systems', Yokohama, Japan, pp. 433–440.

MacLennan, B. J. (1990), Evolution of Communication in a Population of Simple Machines, Technical Report Computer Science Department Technical Report CS–90–104, University of Tennessee.

Matarić, M. J. (1992), Designing Emergent Behaviors: From Local Interactions to Collective Intelligence, *in* J.-A. Meyer, H. Roitblat & S. Wilson, eds, 'From Animals to Animats: International Conference on Simulation of Adaptive Behavior'.

Matarić, M. J. (1994*a*), Interaction and Intelligent Behavior, Technical Report AI-TR-1495, MIT Artificial Intelligence Lab.

Matarić, M. J. (1994*b*), Reward Functions for Accelerated Learning, *in* W. W. Cohen & H. Hirsh, eds, 'Proceedings of the Eleventh International Conference on Machine Learning (ML-94)', Morgan Kauffman Publishers, Inc., New Brunswick, New Jersey, pp. 181–189.

Mataric, M. J. & Simsarian, K. T. (1995), Learning to Cooperate Using Two Six–Legged Mobile Robots, *in* 'Submitted to IJCAI-95'.

Matarić, M. J., Nilsson, M. & Simsarian, K. T. (1994), Cooperative Multi–Robot Box–Pushing, *in* 'submitted to IROS-95', Pittsburgh, PA.

McFarland, D. (1985), *Animal Behavior*, Benjamin Cummings.

McFarland, D. (1987), The Oxford Companion to Animal Behavior, *in* 'Oxford, University Press'.

Millán, J. D. R. (1994), Learning Reactive Sequences from Basic Reflexes, *in* 'Proceedings, Simulation of Adaptive Behavior SAB-94', The MIT Press, Brighton, England, pp. 266–274.

Minsky, M. L. (1954), Theory of Neural–Analog Reinforcement Systems and Its Application to the Brain–Model Problem, PhD thesis, Princeton.

Pomerleau, D. A. (1992), Neural Network Perception for Mobile Robotic Guidance, PhD thesis, Carnegie Mellon University, School of Computer Science.

Samuel, A. L. (1959), 'Some studies in machine learning using the game of checkers', *IBM Journal of Research and Development* **3**, 211–229.

Smithers, T. (1994), On Why Better Robots Make it Harder, *in* 'The Third International Conference on Simulation of Adaptive Behavior'.

Steels, L. (1994), Emergent Functionality of Robot Behavior Through On-line Evolution, *in* R. Brooks & P. Maes, eds, 'Artificial Life IV, Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems', The MIT Press.

Sutton, R. (1988), 'Learning to Predict by Method of Temporal Differences', *The Journal of Machine Learning* **3**(1), 9–44.

Travers, M. (1988), Animal Construction Kits, *in* C. Langton, ed., 'Artificial Life', Addison–Wesley.

Werner, G. M. & Dyer, M. G. (1990), Evolution of Communication in Artificial Organisms, Technical Report UCLA–AI–90–06, University of California, Los Angeles.