Symposium on Applications of Reinforcement Learning

Final Report for NSF Grant IIS-9810208

Pat Langley and Mark Pendrith
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

The Symposium on Applications of Reinforcement Learning was held at Stanford University on March 21 and 22, 1998. Nearly 30 researchers attended the meeting, of which 14 participants presented talks on their results in this area. The symposium fostered a "working atmosphere" in which scientists from different paradigms that would not typically interact could meet to share views and approaches to challenging issues.

Research on reinforcement learning deals with the task of improving an agent's behavior through interaction with an uncertain environment. Thus, this field is more naturally characterized as a class of *problems* than as a set of *methods*. One broad class of techniques involves learning a utility value function for states and/or state-action pairs the reinforcement learning agent will encounter; this includes methods like Q learning, which Sutton and Barto (1998) have described at length. Another generic approach involves carrying out search through the space of control policies; one common scheme relies on genetic search, as Moriarty, Schultz, and Grefenstette (1997) recount. One goal of this symposium was to foster greater communication between these traditionally separate communities within the broader field of machine learning.

Another important objective of the meeting was to review, consolidate, and stimulate new work in the application of reinforcement learning. This field has recently made important advances in theoretical understanding, particularly in the area of value-function techniques, and yet the number of real-world applications has remained relatively small. However, there have been a number of notable successes, and these provided an important focus for the symposium. In general, we need a better understanding of the characteristics of the problem domains that are well-suited to reinforcement learning methods. We hoped the symposium would help resolve some of these issues and lead to the identification of others.

Participation in the symposium was by invitation only. The invited speakers reported on their work over two days, interleaved by extended breaks to encourage informal discussion. We requested that speakers focus their presentations not on their learning algorithms, as in typical academic talks, but on the application problem and the design decisions that produced successful results. In particular, we asked them to discuss how they formulated their application problem in terms of learning from delayed reward, the features they used to describe states and control policies, the reward or fitness function they utilized, and how they evaluated their system's behavior.

We organized the symposium into 30-minute talks, followed by 15-minute discussions, on two consecutive days, interleaving talks from two theoretical frameworks to emphasize common issues rather than paradigmatic differences. We concluded each day with an extended discussion period. Below we summarize the contents for each speaker's presentation in the order they were given.

1. Learning Gait Control in a Walking Robot

Mark Pendrith (Daimler-Benz Research & Technology Center) examined a mobile robotics task in order to focus on the practical issues that arise in the application of reinforcement learning. He discussed the interrelated issues of problem formulation, representation engineering, and defining the reward function. Although his "learning to walk" application was relatively modest in scale, it raised issues that are common to many non-trivial domains. An interesting phenomenon, which Pendrith called *subversive reinforcement learning*, arises when the developer thinks about an agent's desired behavior in terms of tasks or goals, but specifies it only indirectly in terms of a reward function. He proposed that the problem of subversion points to a distinct but less well-recognized class of scaling problem for reinforcement learning.

2. Evolving Complex Robotic Behaviors

Alan Schultz (Naval Research Laboratory) reported progress on a number of problems related to multi-robot scenarios, including collision avoidance, navigation from one location to another, and using some robots to 'herd' other robots in a certain direction. Such tasks lend themselves to solutions that involve reactive controllers, but developing such controllers by hand is a tedious and uncertain process. Instead, Schultz has used reinforcement learning to acquire controllers for these tasks from experience. In particular, he described his work with SAMUEL, a system that relies on evolutionary algorithms to generate alternative robot behaviors and select promising ones based on their performance. To speed the process, Schultz let the system evaluate different controllers on a mobile robotics simulation, then transferred the best learned strategy to his Nomad 200 robots. His successful results suggest that simulators can play an important role in the application of reinforcement learning to practical problems.

3. Successful Learning in Distributed Multi-Robot Systems

Maja Mataric (University of Southern California) described her approach to using 'behaviors' as a representation that enables groups of physical mobile robots to improve their individual and group responses in real time, over the period of 15 to 30 minutes. This approach employs an adaptation of reinforcement learning applied to a behavior-based substrate, enabling improved individual and group efficiency, as well as automated task division and specialization within the group, without external input from the user. This framework has supported learning in groups of mobile robots that are engaged in a variety of tasks and situated in noisy, non-stationary conditions, thus opening the door for adaptive multi-robot applications capable of adjusting themselves automatically to the changing dynamics of their task and environment.

4. Fuzzy Reinforcement Learning for the Space Shuttle Training Aircraft

Hamid Berenji (NASA Ames Research Center) reported an application involving the Space Shuttle Training Aircraft, which NASA uses to train astronauts to land the Shuttle after returning from Earth orbit. He has used fuzzy reinforcement learning to refine rule-based controllers for Space Shuttle landing and thus reduce trajectory-following errors. The new system includes a feed-forward controller, which provides an elevator command to track the pitch rate trajectory, and a learned feedback controller, which eliminates most of the remaining errors. Berenji presented the results of testing this new control system on the Shuttle ground simulator. The new hybrid controller significantly improves the accuracy of trajectory following.

5. Elevator Dispatching Using Multiple Reinforcement Learning Agents

Andrew Barto (University of Massachusetts, Amherst) addressed the problem of efficiently dispatching elevators in multi-story buildings, to which he and his colleagues have applied several multi-agent reinforcement learning methods. The utilized a discrete-event simulation of a ten-story building with four elevator cars and stochastic passenger arrivals, with a separate reinforcement learning agent responsible for controlling each elevator car. Each agent employed a variant of Q learning for discrete-event systems and used feedforward networks to store the Q values. They compared an architecture in which the agents shared the same Q values with one that maintained these values separately. Experimental studies showed performance superior to the best known heuristic elevator control algorithms. They also experimented with variations in state representation, reward computation, and other design decisions.

6. Automobile Traffic Management through Intelligent Lane Selection

David Moriarty (Information Sciences Institute, USC) described joint work with Pat Langley (Daimler-Benz) on a novel approach to traffic management through coordinating driver behaviors. Current traffic management systems do not always consider lane organization of the cars and often only affect traffic flows by controlling traffic signals or ramp meters. However, drivers could increase traffic throughput and maintain desired speeds more consistently by selecting lanes intelligently. Moriarty posed the problem of intelligent lane selection as a challenging and potentially rewarding problem for artificial intelligence, and he proposed a methodology that uses supervised and reinforcement learning to form distributed control strategies. He presented promising initial results demonstrating that learned intelligent lane selection can achieve higher traffic throughput, maximize desired speeds, and reduce the total number of lane changes.

7. Dynamic Channel Allocation in Cellular Telephone Systems

Satinder Singh (University of Colorado, Boulder) discussed the dynamic allocation of communication resource (channels) in cellular telephone systems so as to maximize service in a stochastic caller environment. He formulated this problem as a dynamic programming task and used reinforcement learning to find dynamic channel allocation policies that are better than previous solutions. The learned policies perform well for a broad variety of call traffic patterns. Singh presented results on a large cellular system with approximately 49^{49} states.

8. Call Admission Control and Routing in Integrated Service Networks

John Tsitsiklis (Massachusetts Institute of Technology) discussed the problem of optimal admission control and routing in an integrated service network. He noted that, although one can formulate this task as a dynamic programming problem, it is too complex to be solved exactly. Tsitsiklis and his co-workers instead used reinforcement learning methods to train a cost-to-go function, which in turns results in a dynamic admission control and routing policy. This includes standard discounted TD(λ), as well as a relatively new variant of the method for average-cost problems. They experimented with a 4-node and a 16-node network, using decomposable function approximation to handle the many state variables. Performance of the learned controller compared favorably to that of a commonly used handcrafted policy.

9. Backgammon, Co-Evolution, and the Meta-game of Learning

Alan Blair (University of Queensland) described joint work with Jordan Pollack (Brandeis University) on a simple hill-climbing algorithm that achieves results similar to those reported for Gerald Tesauro's initial TD-Gammon system. Applying reinforcement learning to strategic games generally involves some form of co-evolution, which adds a twist to the exploration-exploitation trade-off because exploration depends on the opponent's actions as well as one's own. Co-evolutionary learning can be modeled as a kind of meta-level game between abstract entities that he called the performer, infiltrator, and evaluator. Stable, suboptimal solutions appear as Nash equilibria in this meta-game. Blair discussed how certain attributes of backgammon and other domains work to prevent such suboptimal equilibria. A better understanding of these issues may lead to improved design of future systems that incorporate co-evolutionary learning.

10. Genetic Programming of Near Minimum Time Spacecraft Attitude Maneuvers

Brian Howley (Lockheed Martin Missiles & Space Company) addressed the task of determining large-angle maneuvers for spacecraft, which is a nonlinear multi-dimensional problem. Solutions based on calculus require iterative calculations that are poorly suited for real-time implementation. Instead, typical spacecraft attitude control systems follow trajectories that satisify actuator constraints but are not time optimal. Howley described how he used genetic programming methods to create attitude control laws that outperform traditional controllers, acheiving maneuver times within two percent of optimal. He discussed his decisions about problem formulation and representation, as well as limitations of genetic programming that he encountered.

11. Simulation-Based Optimization of Lean Manufacturing Systems

Sridhar Mahadevan (Michigan State University) described joint work with Nick Marchalleck and Georgios Theocharous on applications of simulation-based reinforcement learning to optimize flexible manufacturing systems. They focused on optimizing transfer lines in industrial production, which are a sequence of flexible machines isolated by product buffers. A desirable goal of a lean manufacturing system is to maximize satisfied demand while minimizing work-in-process inventory. Mahadevan described a model-free average-reward semi-Markov algorithm for optimizing a transfer line, implemented on a commercial discrete-event simulation software package. Their system, combined with a neural network function approximator, appears to outperform classic widely-used industrial heuristics for optimizing transfer lines. He concluded by assessing the overall framework of discrete-event Markov models, which appears to be very promising, both in terms of its theoretical generality and its practical applicability to manufacturing and other domains.

12. Value Function Approximation for Production Scheduling

Justin Boyan (Carnegie Mellon University) discussed the challenge of production scheduling, a critical problem throughout the manufacturing industry, which involves sequentially configuring a factory to meet forecasted demands. The requirement of maintaining product inventories in the face of unpredictable demand and stochastic output makes standard approaches to scheduling inadequate. Current algorithms, such as simulated annealing and constraint propagation, must employ frequent replanning and other expensive techniques to cope with uncertainty. Boyan presented joint work with Jeff Schneider and Andrew Moore on a Markov decision process formulation of production scheduling which captures stochasticity in both production and demands, using a value

function which can be used to generate scheduling decisions online. He also described an industrial application, along with two reinforcement learning methods for generating an approximate value function in this domain. The results suggests that, in both deterministic and noisy scenarios, value function approximation is an effective technique.

13. Reinforcement Learning for Space Shuttle Payload Processing

Wei Zhang (The Boeing Company) focused on an application of reinforcement learning to payload processing for the Space Shuttle. This involves frequent rescheduling over thousands of tasks with changing requirements, so it would benefit from an approach that is more rapid and adaptive than standard methods. Describing work done jointly with Tom Dietterich (Oregon State University) he reviewed the reinforcement learning methods developed for this application. The key idea involved analyzing a set of training problems and learning a value function to direct search on new problem instances. He discussed issues that arose along the way, including how to formulate the problem in terms of reinforcement learning, which features to use, and how to assess behavior. Zhang presented experimental results which suggested that their approach outperforms previous methods and he discussed the general lessons learned from this endeavor.

14. Learning Object Recognition Strategies

Bruce Draper (Colorado State University) reviewed progress in computer vision over the past 20 years on techniques for feature detection, intermediate-level grouping, depth reconstruction from stereo and/or motion, pose determination, and model matching. However, the number of practical vision systems remains small because these modules are difficult to integrate into functional systems for object recognition. Although the library of vision procedures keeps growing, we still cannot select, parameterize, and sequence algorithms to perform specific tasks. In response, Draper presented an approach to the computational control of vision that addresses this problem. His approach adopts a model in which vision procedures are the actions in a Markov decision process and in which intermediate results (e.g., lines and regions) are the states. The goal is to learn object recognition strategies that dynamically select which procedures to execute, based on the target object class and the results of previous procedures. He reported encouraging results with this approach on some challenging computer vision tasks.

In general, speakers followed the organizers' request that they focus on issues other than their specific learning algorithm, which led to the nearly complete absence of cross-paradigm arguments. Instead, discussions focused on the intended topics of problem formulation, representation engineering, selection of reward/fitness functions, and evaluation, with researchers from both theoretical frameworks contributing very similar comments. One speaker stated that it was refreshing to have an opportunity to discuss openly challenges that every researcher confronts, but that are not considered appropriate material for inclusion in academic papers.

In the process, we believe this open dialogue heightened participants' awareness of applications for learning from delayed reward, improved their understanding of the issues that arise in such applied work, and increased their respect for alternative approaches to this important class of scientific problems. In summary, the Symposium on Applications of Reinforcement Learning achieved its original goals and, we believe, served an important function for the research community at an important time in its development.