Reinforcement Learning:
What We Know, What We Need

An ML93 Workshop

June 30 - July 1

Schedule

June 30:

9:00--10:30	Session 1: Defining Features of Reinforcement Learning
10:30--11:00	Break
11:00--12:30	Session 2: RL and Dynamic Programming
12:30--2:00	Lunch
2:00--3:30	Session 3: Theory: New Results in RL and Asynchronous DP
3:30--4:00	Break
4:00--5:00	Session 4: Hidden State and Short-Term Memory

July 1:

9:00--10:30	Session 5: Structural Generalization: Scaling RL to Large State Spaces
10:30--11:00	Break
11:30--12:30	Session 6: Hierarchy and Abstraction
12:30--1:30	Lunch
1:30--2:20	Session 7: Strategies for Exploration
2:30--3:30	Session 8: Relationships to Neuroscience and Evolution

Sessions

Session 1: Defining Features of Reinforcement Learning
Wednesday, 9:00-10:30
Organizer: Rich Sutton, rich@gte.com

"Welcome and Announcements" by Rich Sutton, GTE (10 minutes)
"History of Reinforcement Learning" by Harry Klopf, WPAFB (25 minutes)
"Delayed Reward: TD Learning" by Rich Sutton, GTE (25 minutes)
"TD-Gammon Achieves Master-Level Play" by Gerry Tesauro, IBM (25 minutes)

The intent of the first two talks is to start getting across certain key ideas about reinforcement learning: 1) reinforcement learning is a problem, not a class of algorithms, 2) the distinguishing features of the RL problem are trial-and-error search and delayed reward. The third talk is a tutorial presentation of temporal-difference learning, the basis of learning methods for handling delayed reward. The last talk will present TD-Gammon, a TD-learning system that learned to play backgammon at a near- grandmaster level.

Session 2: RL and Dynamic Programming
Wednesday, 11:00-12:30
Organizer: Andy Barto, barto@cs.umass.edu

Speakers:
Andy Barto, UMass
Andrew Moore, MIT
Chris Watkins, Morning Side Inc

These talks will cover the basic ideas of reinforcement learning and its relationship to dynamic programming and planning. Markov Decision Tasks will be introduced.

Session 3: Theory: New Results in RL and Asynchronous DP
Wednesday, 2:00-3:30
Organizer: Satinder Singh, singh@cs.umass.edu

"Introduction, Notation, and Theme" by Satinder P. Singh, UMass
"Stochastic Approximation: Convergence Results" by T. Jaakkola & M. Jordan, MIT
"Asychronous Policy Iteration" by Ron Williams, Northeastern
"Convergence Proof of Adaptive Asynchronous DP" by Vijakumar Gullapalli, UMass

This session consists of two parts. In the first part, we present a new and fairly complete theory of (asymptotic) convergence for reinforcement learning (with lookup tables as function approximators). This theory explains RL algorithms as replacing the full-backup operator of classical dynamic programming algorithms by a random backup operator that is unbiased. We present an extension to classical stochastic approximation theory (e.g., Dvoretzky's) to derive probability one convergence proofs for Q-learning, TD(0), and TD(λ), that are different, and perhaps simpler, than previously available proofs. We will also use the stochastic approximation framework to highlight the contribution made by reinforcement learning algorithms such as TD, and Q-learning, to the entire class of iterative methods for solving the Bellman equations associated with Markovian Decision Tasks.

The second part deals with contributions by RL researchers to asynchronous DP. Williams will present a set of algorithms (and convergence results) that are asynchronous at a finer grain than classical asynchronous value iteration, but still use "full" backup operators. These algorithms are related to the modified policy iteration algorithm of Puterman and Shin, as well as to the ACE/ASE (actor-critic) architecture of Barto, Sutton and Anderson. Gullapalli will present a proof of convergence for "adaptive" asynchronous value iteration that shows that in order to ensure convergence with probability one, one has to place constraints on how many model-building steps have to be be performed between two consecutive updates of the value function.

Session 4: Hidden State and Short-Term Memory
Wednesday, 4:00-5:00
Organizer: Lonnie Chrisman, lonnie.chrisman@cs.cmu.edu
Speakers: Lonnie Chrisman & Michael Littman, CMU

Many realistic agents cannot directly observe every relevant aspect of their environment at every moment in time. Such hidden state causes problems for many reinforcement learning algorithms, often causing temporal differencing methods to become unstable and making policies that simply map sensory input to action insufficient.

In this session we will examine the problems of hidden state and of learning how to best organize short-term memory. I will review and compare existing approaches such as those of Whitehead & Ballard, Chrisman, Lin & Mitchell, McCallum, and Ring. I will also give a tutorial on the theories of Partially Observable Markovian Decision Processes, Hidden Markov Models, and related learning algorithms such as Balm-Welsh/EM as they are relevant to reinforcement learning.

Session 5: Structural Generalization: Scaling RL to Large State Spaces
Thursday, 9:00-10:30
Organizer: Sridhar Mahadevan, sridhar@watson.ibm.com

"Motivation and Introduction" by Sridhar Mahadevan, IBM
"Neural Nets" by Long-Ji Lin, Siemens
"CMAC" by Tom Miller, Univ. New Hampshire
"Kd-trees and CART" by Marcos Salganicoff, UPenn
"Learning Teleo-Reactive Trees" by Nils Nilsson, Stanford
"Function Approximation in RL: Issues and Approaches" by Richard Yee, UMass
"RL with Analog State and Action Vectors", Leemon Baird, WPAFB

RL is slow to converge in tasks with high-dimensional continuous state spaces, particularly given sparse rewards. One fundamental issue in scaling RL to such tasks is structural credit assignment, which deals with inferring rewards in novel situations. This problem can be viewed as a supervised learning task, the goal being to learn a function from instances of states, actions, and rewards. Of course, the function cannot be stored exhaustively as a table, and the challenge is devise more compact storage methods. In this session we will discuss some of the different approaches to the structural generalization problem.

Session 6: Hierarchy and Abstraction
Thursday, 11:00-12:30
Organizer: Leslie Kaelbling, lpk@cs.brown.edu

Speakers:
Andrew Moore (MIT)
Peter Dayan (Salk)
Long-ji Lin (Siemens)

Too much of RL is concerned with low-level actions and low-level (single time step) models. How can we model the world, and plan about actions, at a higher level, or over longer time scales? How can we integrate models and actions at different time scales and levels of abstraction? To address these questions, several researchers have proposed models of hierarchical learning and planning, e.g., Satinder Singh, Mark Ring, Chris Watkins, Long-ji Lin, Leslie Kaelbling, and Peter Dayan & Geoff Hinton. The format for this session will be a brief introduction to the problem by the session organizer followed by short talks and discussion.

Session 7: Strategies for Exploration
Thursday, 1:30-2:20
Organizer: Steve Whitehead, swhitehead@gte.com

Exploration is essential to reinforcement learning, since it is through exploration, that an agent learns about its environment. Naive exploration can easily result in intractably slow learning. On the other hand, exploration strategies that are carefully structured or exploit external sources of bias can do much better.

A variety of approaches to exploration have been devised over the last few years (e.g., Kaelbling, Sutton, Thrun, Koenig, Lin, Clouse, Whitehead). The goal of this session is to review these techniques, understand their similarities and differences, understand when and why they work, determine their impact on learning time, and to the extent possible organize them taxonomically.

The session will consist of a short introduction by the session organizer followed by a open discussion. The discussion will be informal but aimed at issues raised during the monologue. An informal panel of researchers will be on hand to participate in the discussion and answer questions about their work in this area.

Session 8: Relationships to Neuroscience and Evolution
Thursday, 2:30-3:30
Organizer: Rich Sutton, rich@gte.com

"Classifier Systems as Reinforcement Learners" by Stewart Wilson, Rowland Institute.
Classifier systems were introduced by Holland in 1976 for learning problems requiring temporal credit allocation, structural generalization, and sensitivity to hidden states--in effect, RL problems. The systems are based on an evolving population of initially random condition/action rules which compete to control the system. The talk will briefly describe the "canonical" classifier system model plus simplified versions which have achieved concrete results. The bucket-brigade's resemblance to Q-learning, fitness measures for generalization, and techniques for representing hidden states and time-scales will be sketched.

"RL in the Brain: Developing Connections Through Prediction" by R Montague, Salk.
Both vertebrates and invertebrate possess diffusely projecting neuromodulatory systems. In the vertebrate, it is known that these systems are involved in the development of cerebral cortical structures and can deliver reward and/or salience signals to the cerebral cortex and other structures to influence learning in the adult. Recent data in primates suggest that this latter influence obtains because changes in firing in nuclei that deliver the neuromodulators reflect the difference in the predicted and actual reward, i.e., a prediction error. This relationship is qualitatively similar to that predicted by Sutton and Barto's classical conditioning theory. These systems innervate large expanses of cortical and subcortical turf through extensive axonal projections that originate in midbrain and basal forebrain nuclei and deliver such compounds as dopamine, serotonin, norepinephrine, and acetylcholine to their targets. The small number of neurons comprising these subcortical nuclei relative to the extent of the territory their axons innervate suggests that the nuclei are reporting scalar signals to their target structures. These facts are synthesized into a single framework which relates the development of brain structures and conditioning in adult brains. We postulate a modification to Hebbian accounts of self-organization: Hebbian learning is conditional on a incorrect prediction of future delivered reinforcement from a diffuse neuromodulatory system. The reinforcement signal is derived both from externally driven contingencies such as proprioception from eye movements as well as from internal pathways leading from cortical areas to subcortical nuclei. We suggest a specific model for how such predictions are made in the vertebrate and invertebrate brain. We illustrate the framework with examples ranging from the development of sensory and sensory-motor maps to foraging behavior in bumble-bees.