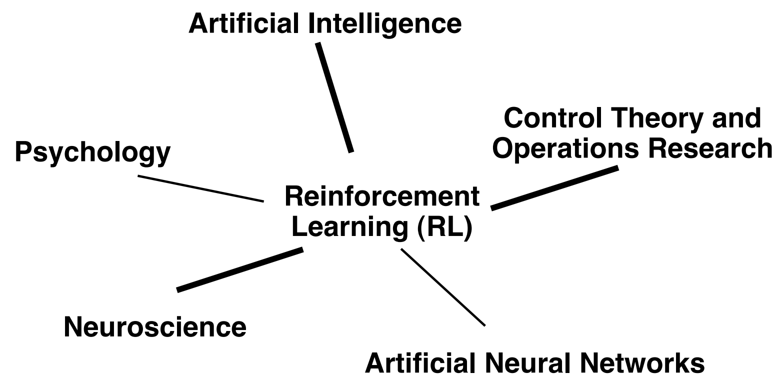
Computer Science 687
Spring 2006

Reinforcement Learning

Andrew Barto
Department of Computer Science
University of Massachusetts — Amherst

Lecture 1: Introduction

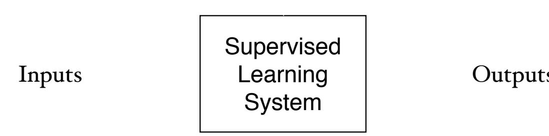


What is Reinforcement Learning?

- Learning from interaction
- Goal-oriented learning
- Learning about, from, and while interacting with an external environment
- Learning what to do—how to map situations to actions —so as to maximize a numerical reward signal

Supervised Learning

Training Info = desired (target) outputs



Error = (target output − actual output)

Reinforcement Learning

Training Info = evaluations (“rewards” / “penalties”)

Inputs → **Reinforcement Learning System** → Outputs (“actions”)

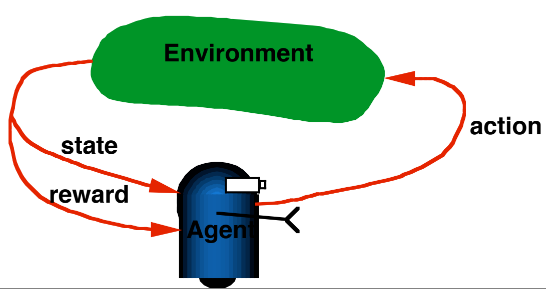
Objective: get as much reward as possible

Key Features of RL

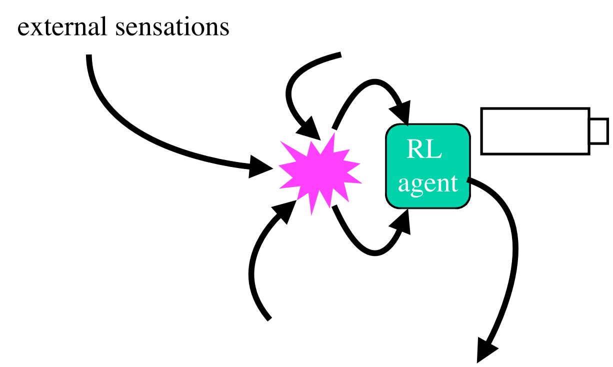
- Learner is not told which actions to take
- Trial-and-Error search
- Possibility of delayed reward
  - Sacrifice short-term gains for greater long-term gains
- The need to ***explore*** and ***exploit***
- Considers the whole problem of a goal-directed agent interacting with an uncertain environment

Complete Agent

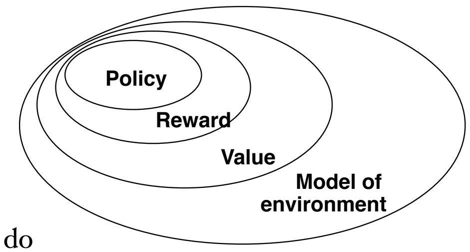
- Temporally situated
- Continual learning and planning
- Object is to ***affect*** the environment
- Environment is stochastic and uncertain



A Less-Misleading Agent View

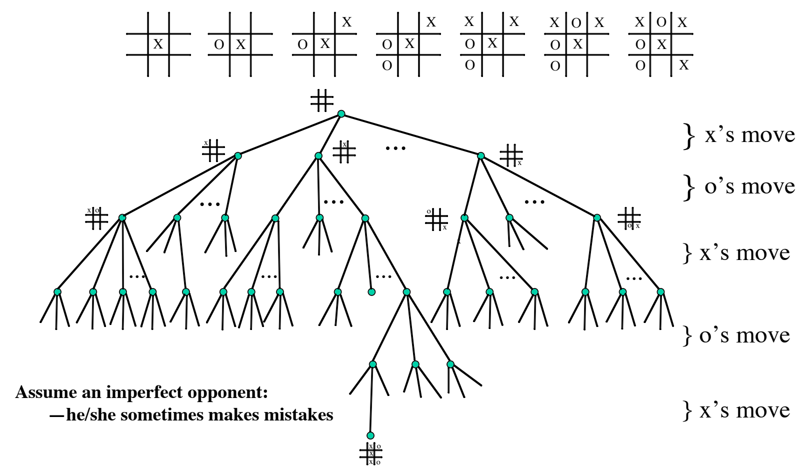


Elements of RL



- **Policy**: what to do
- **Reward**: what is good
- **Value**: what is good because it ***predicts*** reward
- **Model**: what follows what

An Extended Example: Tic-Tac-Toe

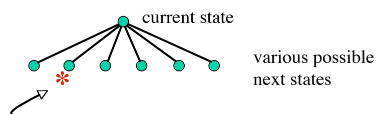


An RL Approach to Tic-Tac-Toe

1. Make a table with one entry per state:

State	$V(s)$ – estimated probability of winning	
	.5	?
	1	win
	0	loss
	0	draw

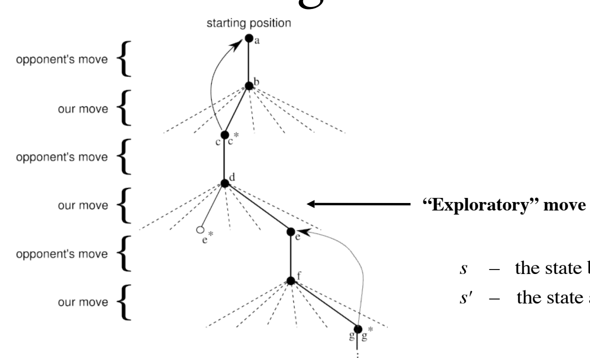
2. Now play lots of games. To pick our moves, look ahead one step:



Just pick the next state with the highest estimated prob. of winning — the largest $V(s)$; a ***greedy*** move.

But 10% of the time pick a move at random; an ***exploratory move***.

RL Learning Rule for Tic-Tac-Toe



s – the state before our greedy move
s' – the state after our greedy move

We increment each $V(s)$ toward $V(s')$ – a ***backup***:

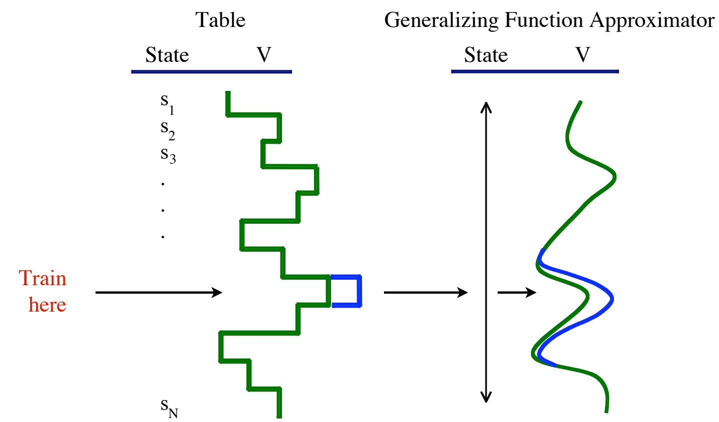
$$V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$$

a small positive fraction, e.g., $\alpha = .1$
the ***step - size parameter***

How can we improve this T.T.T. player?

- Take advantage of symmetries
  - representation/generalization
  - How might this backfire?
- Do we need "random" moves? Why?
  - Do we always need a full 10%?
- Can we learn from "random" moves?
- Can we learn offline?
  - Pre-training from self play?
  - Using learned models of opponent?
- . . .

e.g. Generalization



How is Tic-Tac-Toe Too Easy?

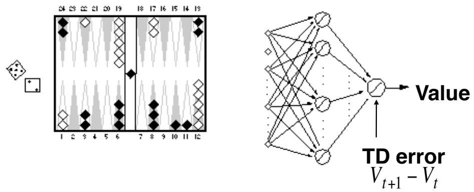
- Finite, small number of states
- One-step look-ahead is always possible
- State completely observable
- . . .

Some Notable RL Applications

- **TD-Gammon**: Tesauro
  - world's best backgammon program
- **Elevator Control**: Crites & Barto
  - high performance down-peak elevator controller
- **Inventory Management**: Van Roy, Bertsekas, Lee & Tsitsiklis
  - 10–15% improvement over industry standard methods
- **Dynamic Channel Assignment**: Singh & Bertsekas, Nie & Haykin
  - high performance assignment of radio channels to mobile telephone calls
- **More ...**

TD-Gammon

Tesauro, 1992–1995



Action selection by 2–3 ply search

Start with a random network

Play very many games against self

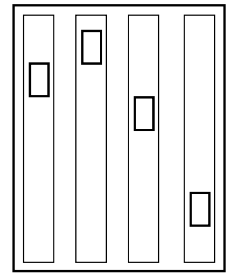
Learn a value function from this simulated experience

This produces arguably the best player in the world

Elevator Dispatching

Crites and Barto, 1996

10 floors, 4 elevator cars



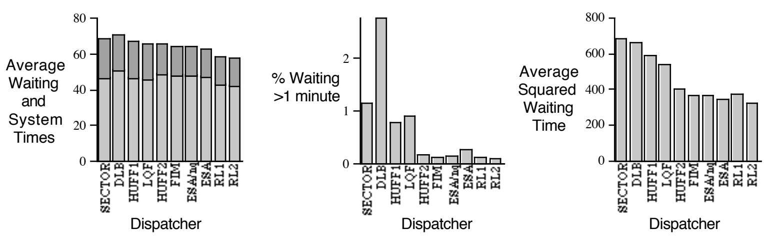
STATES: button states; positions, directions, and motion states of cars; passengers in cars & in halls

ACTIONS: stop at, or go by, next floor

REWARDS: roughly, –1 per time step for each person waiting

Conservatively about 10^{22} states

Performance Comparison



Autonomous Helicopter Flight

A. Ng, Stanford, H. Kim, M. Jordan, S. Sastry, Berkeley

Quadrupedal Locomotion

Nate Kohl & Peter Stone, Univ of Texas at Austin

All training done with physical robots: Sony Aibo ERS-210A

Before learning

After 1000 trials, or about 3 hours

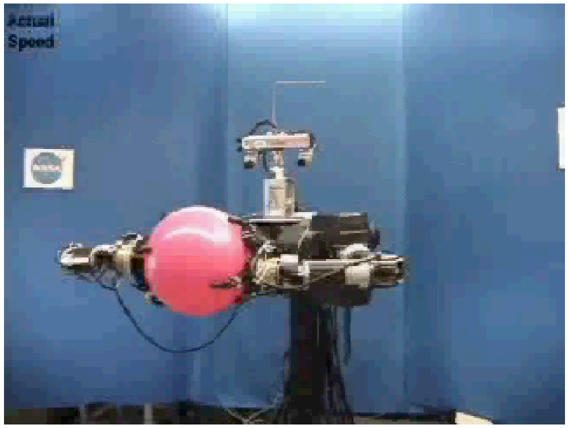
Learning Control for Dynamically Stable Walking Robots

Russ Tedrake, Teresa Zhang, H. Sebastian Seung, MIT

http://hebb.mit.edu/~russt/robots

Grasp Control

R. Platt, A. Fagg, R. Grupen, Univ. of Mass Amherst

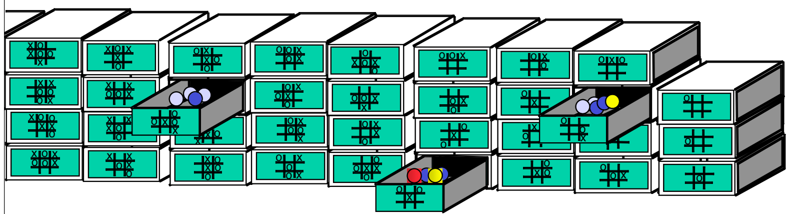


Some RL History

Trial-and-Error learning	Temporal-difference learning	Optimal control, value functions
Thorndike (Ψ) 1911	Secondary reinforcement (Ψ)	Hamilton (Physics) 1800s
		Shannon
Minsky	Samuel	Bellman/Howard (OR)
Klopf	Holland	
	Witten	Werbos
Barto et al.		
	Sutton	
		Watkins

MENACE (Michie 1961)

"Matchbox Educable Noughts and Crosses Engine"



The Book

- Part I: The Problem
  - Introduction
  - Evaluative Feedback
  - The Reinforcement Learning Problem
- Part II: Elementary Solution Methods
  - Dynamic Programming
  - Monte Carlo Methods
  - Temporal Difference Learning
- Part III: A Unified View
  - Eligibility Traces
  - Generalization and Function Approximation
  - Planning and Learning
  - Dimensions of Reinforcement Learning
  - Case Studies

The Course

- We will follow the book, then read a collection of more recent papers on later developments
- Read the reading assignment for the each class before that class!
- Written home-works: many of the non-programming assignments in each chapter, plus others.
- Programming exercises: require you to implement many of the algorithms discussed in the book. Details to come...
- Closed-book, in-class midterm; closed-book 2-hr final
- Grading: 30% written home-works; 25% programming exercises; 25% final; 20% midterm
- See the web for more details: http://www-anw.cs.umass.edu/~barto/courses/cs687/

Next Class

- Introduction continued and some case studies
- Read Chapters 1 & 2
- Do exercises 1.1 — 1.5: to hand in Tues 2/7