Sutton & Barto

Ch 1

1.1
1.2
1.3

- "the central role of value estimation is arguably the most important thing that has been learned about RL over the last six decades"
- in ch 8 we explore RL systems that simultaneously learn by trial and error, learn a model of the environment, and use the model for planning (*cool*)
  1.4
- most of the RL systems in the book focus on estimating value functions
- RL systems that don't estimate value functions: genetic algos, genetic programming, simulated annealing, other optimization methods (these are all *evolutionary* methods)
  - don't learn by interacting with env
  - useful when policy space is small, or good policies are easy to find
  - useful in inaccessible envs
    1.6
- Bellman extended work by Hamilton and Jacobi to form Bellman equation
- class of methods to solve optimal control problems by solving this equation called *dynamic programming*
- discrete stochastic version of optimal control problem called **MDP**
- Ronald Howard devised policy iteration method for MPDs.
  1.7
- Edward Thorndike's *Law of Effect*: describes effect of reinforcing events on the tendancy to select actions
  - controversial across disciplines
- see cyberneticzoo.com for early trial-and-error learning machines
- must read Minsky's paper "Steps Toward Artificial Intelligence" (1961)
  - *basic credit-assignment problem for complex reinforcement learning systems*: How do you distribute credit for success among the many decisions that may have been involved in producing it?
  - all methods in this book essential attempt to solve that problem
- more readings

- John Andreae, STeLLA system: internal model of world and "internal monologue" to deal with hidden state, goal of generating novel events
- Donal Michie, MENACE: tic-tac-toe
- Widrow, Gupta, Maitra modified Least-Mean-Square (LMS) algo to produce rule to learn from success and failure signals instead of training samples, our term *critic* originates here
- ***learning automata*: methods for solving nonassociative, purely selectional learning problems like multi-arm bandit**
- statistical learning theories applied to econ -> incorporation of game theory
- John Holland (1975) - general theory of adaptive systems based on selectional principles (*classifier systems*)
- Harry Klopf - most important person for reviving trial-and-error of RL

- temproal-difference learning (TD)
  - *secondary reinforcer*: a stimulus that has been parired with a primary reinforcer such as food or pain and , as a result, has come to take on similar reinforcing properties
  - influenced by animal learning theories
  - Sutton (1988) introduced $TD(\lambda)$ algo and proved some of its convergence properties
- Q-learning (Chris Watkins 1989)
  - TD and optimal control threads brought together

Ch 3

- in bandits we estimate value of $q_*(a)$ of each action a, in MDPs we estimate the avlue of $q_*(s, a)$ of each action in each state or $v_*(s)$ of each state given optimal action
- the added state requirement in MDPs allows credit-assingment for long-term consequences
  3.1
- practice problems and examples of representing problems as MDPs
  3.2
  3.3
  3.4
  3.5
- *state-value function for policy* π : $v_\pi(s) \doteq \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s]]$
- *action-value function for policy* π : $q_\pi(s, a) \doteq \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a]]$
  3.6
- Gridworld example and optimal policy derivation
  3.7
  3.8

Ch 16

16.1

16.2

16.3

16.4

16.5

16.6

16.7

16.8