Littman 1996

Chapter 1

Sequential Decision Making

- agent : system responsible for interacting with the world and making decisions
  - transition function T
  - perception function O (if O is the identify function, agent sees true state of environment)
  - behavior function B maps perception to action choice
  - reward function R
- environment : anything external to agent
  - probabilities output by agent's T must remain constant over time
  - i.e. the laws of the environment don't change (**stationary**)
  - elements external to the agent that follow changing rules can sometimes be thought of as other agents
- reward : specifies the problem the agent is to solve in interactions with environment
  - agents must have subjective access to rewards
  - *cool: Can rewards exist in "undesigned" environments? In bio, true reward signal is death, which happens too late for agent to modify actions. But, simiulations have shown that over generations, agents can evolve their own proximal reward functions.*
- policy : agent's prescription for behavior
  - *plan* is a simple kind of policy in which the agent takes a sequence of actions "with its eyes closed," useful in predictable envs with known inintial state
  - *conditional plan* admits small variation in sequence
  - most of thetime "policy" = "stationary policy" -- the extreme end of conditional plan is the *universal plan* or *stationary policy* in which the agent takes entire state at each step and makes optimal action for current state

Formal Models

- finite vs. continuous state
- finite vs. continous actions
- episodic vs. sequential
- accessible vs. inaccessible (partially observable envs are considered inaccessible)
- Markovian vs. non-Markovian (in M, future evolution of the system can be predicted on the basis of the env's state)
  note: some problems seem non-Markovian simply because they are inaccessible

- fixed vs. dynamic (can the env change while agent is considering an action?)
- deterministic vs. stochastic
- synchronous vs. asynchronous
- single vs. multiple agent

Evaluation Criteria

- policies
  - *objective function* : takes a set of possible state and action sequences and their probabilities and returns a value
  - the optimal policy maximizes the objective function
  - this is like the utility function from the lectures?
    - $U(s) = E[\sum_t \gamma^t R(s_t) | s_0 = s]$ calculated with the **Bellman equation** $U(s) = R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U(s')$
    - recall $E[.]$ represents a probability-weighted mean of possible values of a random variable
  - see pg. 17 for popular objective functions
  - note γ dictacts how much a reward is worth in the future, i.e., a reward r received t steps in the futures is worth $\gamma^t r$ (Littman uses β here)
- planning algos
- RL algos
  - truly optimal RL agent maximizes objective function under the restriction that its knowledge of the env is incomplete (believed to be untractable except in trivial case) **is this still true?**
    - from Farrukh Rahman @23:

> "Yeah when there is only 1 state it can be considered a multi-armed bandit. With a single state we want to select the action (aka pull the arm) that gives us the best expected reward, however we may not have any a priori knowledge of which action to take.
> Two natural methods emerge
>
> 1. take each action many times and build an expectation of what each actions reward is. this would be exploration
> 2. use our knowledge so far and take the action that has the best looking data (ie. maximize likelihood) which is exploitation. This knowledge may be incomplete eg. we only took each action once
>
> There is opportunity cost/regret in each action we take. ie. we cant explore all the time because we lose expected reward, neither can we exploit because we may not have enough confidence in our estimates. We want to balance exploration vs exploitation to obtain the best action while still performing as well as we possible can.
> The Gittins theorem/index incorporates this information (how much we know about

> an arm (how well we've sampled it) vs the maximum likelihood) into a scalar value we can use to do this. ie. we are doing as well as we can while still learning the best action to take.
> In regards to generalizing this outside bandits (>1 states), I do not believe that it the case. Moreover in lectures Littman advises to be cautious with this question as researchers have spent considerable effort trying to generalize it unsuccessfully (Lectures were done in 2015, perhaps some breakthroughs have been made since then but I'm not aware of any).
> Bandits & Gittins will be covered in the course (lesson 8), and are also covered in Sutton Chpt2.

- *regret* : amount of additional reward an algo would have received if it had used the optimal policy from the start
- "good" algos converge to optimal policy, "bad" don't -- comparing rate of convergence is hard

Thesis Summary

The following models are explored and improved in this thesis:

- MDP (new analysis of complexity of policy-iteration algo - see notes)
- Generalized MDP (new model and new proof of convergence of RL in this model)
- Alternating Markov games (first convergence proof for RL)
- Markov games
- Information-state MDP