COMP 652: Machine Learning

Lecture 21

Today

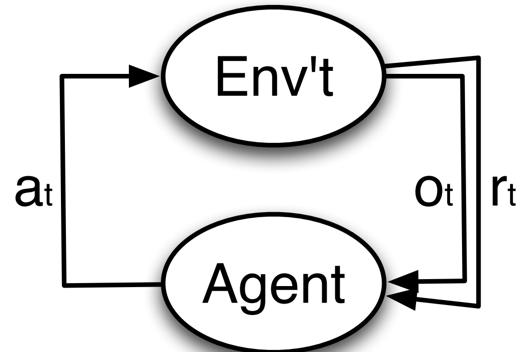
- Partially observable Markov decision processes
- Notions of policy
- Finding optimal policies

Partially Observable Markov Decision Processes (POMDPs)

- Model an agent interacting with an environment, without perfect state information
- Examples:
  - Robot localization, but with active control of robot
  - Sequential strategies for disease therapy
  - Dialogue systems
- Optimal policies may include actions whose primary purpose is to collect state information, even at some cost!

POMDPs formally

- The agent take actions and receives observations and rewards form the environment:



- A POMDP is defined by:
  - A finite state set S
  - A finite action set A
  - A finite observation set O
  - Start state probabilities $p_s = P(S_1 = s)$
  - State transition probabilities $p^a_{ss'} = P(S_{t+1} = s' | S_t = s, A_t = a)$
  - Reward distributions, with expectations $r^a_s = E(r_{t+1} | S_t = s, a_t = a)$
  - Observation probabilities, $p_{so} = P(O_t = o | S_t = s)$
- The $p_s, p^a_{ss'}, r^a_s$ define an “underlying MDP”

The goal?

- Given the POMDP parameters, or experience interacting with the POMDP, find an optimal policy.
- What is a policy in this case?
- How do POMDP policy values compare with underlying MDP values?

Remark

- The value of the optimal policy for the underlying MDP may not be achievable while interacting with the POMDP.
- Intuitively clear: you don't have as much info to act on!
- (Example on board)

What is a policy for a POMDP?

- For MDPs, we restricted attention to deterministic policies $\pi : S \mapsto A$, which specify which action to take based on the current state.
- On what should a POMDP policy depend?
  - The most recent observation, o_t?
  - The k most recent observations, $o_{t-k+1}, o_{t-k+2}, \ldots, o_t$?
  - All previous observations, $o_1, o_2, \ldots, o_t$?
  - All previous observations, actions, and rewards, $o_1, a_1, r_2, o_2, a_2, \ldots, o_t$?
  - The belief state: $P(S_t = s | o_1, o_2, \ldots, o_t)$?
- Should POMDP policies be deterministic or stochastic?
- Should POMDP policies have "memory" or be memoryless?

Remarks and questions

- Policies that depend only on o_t may not be optimal.
- Policies that depend on a length-k history of observations may not be optimal.
  - For any k, I can show you a POMDP with a policy better than any achievable with length-k history.
  - What if we allow k to be chosen to depending on the POMDP?
- Given either of the previous two choices, can stochastic policies outperform deterministic policies?
- What about policies that depend on all prior observations?
- What about policies that depend on all prior observations and actions?
- What about policies that depend on all prior observations, actions and rewards?
- What about policies that depend on belief state?
- What about policies that depend on their own, internal state?

Some answers (with examples done on board)

- Even if k is chosen based on the POMDP, policies depending on length-k history may not be optimal.
- Given either of the previous two choices, can stochastic policies outperform deterministic policies? **Yes.**
- What about policies that depend on all prior observations? **Not optimal in general.**
- What about policies that depend on all prior observations and actions? **Not optimal in general–though often taken to be so.**
- What about policies that depend on all prior observations, actions and rewards? **Includes optimal policy! (By equivalence to a derived MDP.)**
- What about policies that depend on belief state?**Includes optimal policy! Because it's an MDP.**
- What about policies that depend on their own, internal state? **Not optimal in general–may not have enough internal states!**

Sufficient statistics

- Let the history up to time t be $h_t = (o_1, a_1, r_2, o_2, a_2, r_2, \ldots, o_t$.
- Given any action a, there is a well-defined probability:

$$P(r_{t+1} = r, O_{t+1} = o | h_t, a_t = a)$$

- In some cases, not all the information h_t is relevant, however. Or, the information may be summarized in another way.
- In general, let f be some function of history. E.g.:
  - $f(h_t) = h_t$
  - $f(h_t) = o_t$
  - $f(h_t) = (o_{t-k+1}, o_{t-k+2}, \ldots, o_t)$
  - $f(h_t) = (P(s_t = 1|h_t), P(s_t = 2|h_t), \ldots, p(S_t = m|h_t))$, where $S = \{1, 2, \ldots, m\}$.
- f is called a *sufficient statistic* if for all possible histories h_t:

$$P(r_{t+1} + r, O_{t+1} = o | h_t, a_t = a) = P(r_{t+1} + r, O_{t+1} = o | f(h_t), a_t = a)$$

Sufficient statistics (II)

- Suppose f is a sufficient statistic for a given POMDP
- f defines a Markov decision process in the following way:
  - $S = \{f(h_t) : h_t \text{ is a possible history }\}$
  - $A =$ same actions as in the POMDP
  - State transitions $p_{ff'}$: Let $f = f(h_t)$, where $h_t = (o_1, a_1, r_2, \dots, o_t)$. Let $f' = f((o_1, a_1, r_2, \dots, o_t, a_t, r_{t+1}, o_{t+1}))$. Then:

$$p_{ff'} = P(r_{t+1}, o_{t+1} | h_t, a_t)$$

  - Expected rewards: $r^a_{ff'}$ defined similarly.
- Many approaches to solving POMDPs rely on an f which is or is hoped to be a sufficient statistic.
- If it is, then standard approaches for solving/reinforcement learning in MDPs can be applied. (Though S may be infinite now.)
- If it is not, then the behavior of such solution methods cannot be guaranteed in general.

Strategies for finding optimal policies

- Because we don't always have/want to use a sufficient statistic f, and because there are so many notions of policy, there are many different approaches to finding/learning optimal policies.
- (Note that we still haven't defined optimality yet.)
- Solutions also depend on:
  - Whether or not we know the POMDP parameters.
  - Whether we're seeking a deterministic or stochastic policy.
  - Whether we're seeking memoryless policies, or ones with internal state.
  - Whether or not we're using a sufficient statistic, f.

Model-free value-free approaches

- Suppose we don't know the POMDP parameters
- Suppose we have a finite set of policies we want to consider, $\pi_1, \pi_2, \ldots, \pi_m$
- Suppose we can generate sample returns $r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots + \gamma^T r_{t+1}$ under any policy
- We can treat it is a multi-armed bandit problem!

Model-free value-free approaches (II)

- When the multi-armed bandit approach is infeasible, we may perform a more limited, usually local, search in some space of possible policies.
  - E.g., for memoryless, deterministic policies $\pi : O \mapsto A$, there are $|A|^{|O|}$ possible policies. Neighboring policies might different by action assigned to just one observation. Stochastic local search would start with some initial policy, and would repeatedly take samples of neighboring policies until one is found that is better than the current policy.
  - E.g., for stochastic policies, $\pi : O \times A \mapsto [0, 1]$, stochastic gradient descent starts at a policy, and repeatedly takes samples to estimate the gradient of the return w.r.t. the probabilities, and takes a small step in that direction.
  - Similarly for policies that depend on a history of observations, or an internal state.

Value function-based approaches

- Alternatively, we can attempt to estimate a value function V or action-value function Q, treating observations (or sequences of observations and actions) as if they were state:
  - $Q(o, a)$ (depending just on most recent observation)
  - $Q(o_{t-k+1}, a_{t-k+1}, \ldots, o_t, a)$
  - $Q(o_1, a_1, r_1, \ldots, o_t)$ (depending on full history)
- If the observations/actions are a *sufficient statistic* (i.e. equivalent to keeping full history) or nearly so, then Monte Carlo policy evaluation & iteration or Q-Learning can succeed
- If not, then such algorithms may not converge at all, or may converge to something suboptimal (even within the range of policies expressible)
- Relatedly, one can define a feature mapping ϕ that maps each possible history $h_t = (o_1, a_1, r_1, o_2, a_2, r_2, \ldots, o_t)$ to an n-dimensional feature vector, and attempt value function approximation.

Solutions based on belief state

- Suppose we know the POMDP parameters
- The *belief state* is the distribution over possible states, conditioned on observatons and actions so far:

$$b_t(s) = P(S_t = s | o_1, a_1, o_2, a_2, \ldots, o_{t-1}, a_{t-1}, o_t)$$

- Technically, the rewards might be useful in estimating state as well, but they are usually ignored.
- Belief state can be updated easily, based on subsequent observations and actions a_t, o_{t+1}:

$$b_{t+1}(s) = \frac{p_{so_{t+1}} \sum_{s'} p_{s's}^{a_t} b_t(s')}{\sum_{s''} (p_{s''o_{t+1}} \sum_{s'} p_{s's''}^{a_t} b_t(s'))}$$

- Transitions between belief states, depending on actions, along with expected rewards, define a continuous-state MDP.
- There are many ways of computing / approximating such value functions. (They're piecewise linear.)

POMDP summary

- POMDPs model sequential decision-making for an agent interacting with an environment on the basis of imperfect state information.
- Optimal policies need to depend on full observation-action-reward histories, belief state or some other sufficient statistic (and may then be deterministic).
  - In practice, however, this is computationally intensive or outright infeasible.
- Various limited-dependence policies are possible, including ones depending on finite history or memory.
  - Often, one uses a direct search in policy space—the only difference compared to most optimization problems being that we sample the objective function, we do not know it exactly. (Also, *tons* of sample trajectories are typically needed in this approach.)
  - Alternatively, one can hope that a near-enough approximation of state is achieved by your representation, and apply value-function approximation methods