CSC 449/549 Advanced Topics in Artificial Intelligence

Deep Reinforcement Learning

Final Exam

Fall, 2022

Name: ______________________ ID#: ______________ Score: ______

1. (10 pts) Monte Carlo methods for learning value functions require episodic tasks. Why, specifically? How is it that *n*-step TD methods avoid this limitation and can work with continuing tasks?

2. (20 pts) Your Monte-Carlo algorithm generates the following episode using policy π when interacting with its environment. This is the first episode that has been generated.

Timestep	Reward	State	Action
0		s_1	a_1
1	13	s_1	a_2
2	7	s_1	a_1
3	13	s_1	a_2
4	14	s_2	

Assume the discount factor, γ, is 1, and s_2 is a terminal state.

a) What are the estimates of: $q_\pi(s_1, a_1)$ and $q_\pi(s_1, a_2)$ if using first-visit?

b) What are the estimates of: $q_\pi(s_1, a_1)$ and $q_\pi(s_1, a_2)$ if using every-visit?

3. (4 pts) True or False?

a) _____ Q-learning can learn the optimal Q-function Q^* without ever executing the optimal policy.

b) _____ If an MDP has a transition model T that assigns non-zero probability for all triples $T(s, a, s')$ then Q-learning will fail.

4. (16 pts) What is the formal definition of a Partially Observable Markov Decision Process (POMDP), and why is it so much more difficult to find an optimal policy for a POMDP compared to a Completey Obesrvable Markov Decision process?

5. (50 pts) A rat is involved in an experiment. It experiences one episode. At the first step it hears a bell. At the second step it sees a light. At the third step it both hears a bell and sees a light. It then receives some food, worth $+1$ reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted.

a) (7 pts) Represent the rat's state s by a vector of two binary features, $bell(s) \in \{0,1\}$ and $light(s) \in \{0,1\}$. Write down the sequence of feature vectors corresponding to this episode.

b) (7 pts) Approximate the state-value function by a linear combination of these features with two parameters: $b \cdot bell(s) + l \cdot light(s)$. If $b = 2$ and $l = -2$ then write down the sequence of approximate values corresponding to this episode.

c) (4 pts) Define the λ-return v_t^λ.

d) (7 pts) Write down the sequence of λ-returns v_t^λ corresponding to this episode, for $\lambda = 0.5$ and $b = 2$, $l = -2$.

e) (7 pts) Using the forward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? What is the total update to weight b? Use $\lambda = 0.5$, $\gamma = 1$, $\alpha = 0.5$ and start with $b = 2$, $l = -2$.

f) (4 pts) Define the TD(λ) accumulating eligibility trace $\mathbf{e_t}$ when using linear value function approximation.

g) (7 pts) Write down the sequence of eligibility traces $\mathbf{e_t}$ corresponding to the bell, using $\lambda = 0.5$ and $\gamma = 1$,

h) (7 pts) Using the backward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? (Use offline updates, i.e. do not actually change your weights, just accumulate your updates). What is the total update to weight b? Use $\lambda = 0.5$, $\gamma = 1$, $\alpha = 0.5$ and start with $b = 2$, $l = -2$.