Midterm Exam

ELEN E6885: Introduction to Reinforcement Learning

Oct. 22nd, 2021

Formula Sheet

1. Bellman expectation equation:

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)(R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a v_\pi(s'))$$

2. Bellman optimality equation:

$$v_*(s) = \max_a q_*(s,a) = \max_a \{R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a v_*(s')\}.$$

3. TD(0) iteration:

$$v(s_t) \leftarrow v(s_t) + \alpha(r_{t+1} + \gamma v(s_{t+1}) - v(s_t)).$$

Problem 1 (**20 Points, 2 Points each**)

True or False. No explanation is needed.

1. Reinforcement learning uses the formal framework of Markov decision processes (MDP) to define the interaction between a learning agent and its environment in terms of states, actions and rewards.

2. MDP instances with small discount factors tend to emphasize near-term rewards.

3. If the only difference between two MDPs is the value of discount factor γ, then they must have the same optimal policy.

4. Both ϵ-greedy policy and Softmax policy will balance between exploration and exploitation.

5. Every finite MDP with bounded rewards and discount factor $\gamma \in [0, 1)$ has a unique optimal policy.

6. In practice, both policy iteration and value iteration are widely used, and it is not clear which, if either, is better in general.

7. Generalized policy iteration (GPI) is a term used to refer to the general idea of letting policy-evaluation and policy improvement processes interact, independent of the granularity and other details of the two processes.

8. Sarsa converges with probability 1 to an optimal policy and action-value function as long as all state–action pairs are visited an infinite number of times and the policy converges in the limit to the greedy policy.

9. In partially observable Markov decision process (POMDP), it is likely that an agent cannot identify its current state. So the best choice is to maintain a probability distribution over the states and actions. and then updates this probability distribution based on its real-time observations.

10. Many off-policy methods utilize importance sampling, a general technique for estimating expected values under one distribution given samples from another.

Problem 2 (**20 Points, 5 Points each**)

Short-answer questions.

1. Consider a 100x100 grid world domain where the agent starts each episode in the bottom-left corner, and the goal is to reach the top-right corner in the least number of steps. To learn an optimal policy to solve this problem you decide on a reward formulation in which the agent receives a reward of +1 on reaching the goal state and 0 for all other transitions. Suppose you try two variants of this reward formulation, (P1), where you use discounted returns with $\gamma \in (0, 1)$, and (P2), where no discounting is used. As a consequence, a good policy can be learnt in (P1) but no learning in (P2), why?

2. Suppose the reinforcement learning player was greedy. Might it always learn to play better, or worse, than a nongreedy player?

3. Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

4. Given a stationary policy, is it possible that if the agent is in the same state at two different time steps, it can choose two different actions? If yes, please provide an example.

Problem 3 (**30 Points**)

Consider a 2-state MDP. The transition probabilities for the two actions a_1 and a_2 are summarized in the following two tables, where the row and column represents from-state and to-state, respectively. For example, the probability of transition from s_1 to s_2 by taking action a_1 is 0.8. The reward function is $\mathcal{R}_{s_1}^{a_1} = \mathcal{R}_{s_1}^{a_2} = 1$ and $\mathcal{R}_{s_2}^{a_1} = \mathcal{R}_{s_2}^{a_2} = 0$. Assume the discount factor $\gamma = 0.5$.

Table 1: Transition probabilities

(a) action a_1

	s_1	s_2
s_1	0.2	0.8
s_2	0.6	0.4

(b) action a_2

	s_1	s_2
s_1	0.6	0.4
s_2	0.2	0.8

1. [10 Pts] Let v_1^*, v_2^* denote the optimal state-value function of state s_1, s_2, respectively. Write down the Bellman optimality equations with resepect to v_1^* and v_2^*.

2. [6 Pts] Based on your answer to the previous question, prove that $v_1^* > v_2^*$.

3. [14 Pts] Based on the fact that $v_1^* > v_2^*$, find out the optimal value function v_1^* and v_2^*. And write down the optimal policy of this MDP.

Problem 5 (30 Points)

Consider an *undiscounted* MDP ($\gamma = 1$) with two non-terminal states A, B and a terminal state C. The transition function and reward function of the MDP are unknown. However, we have observed the following two episodes:

$$A, a_1, -1, A, a_1, +1, A, a_2, +3, C,$$
$$A, a_2, +1, B, a_3, +2, C,$$

where a_1, a_2, a_3 are actions, and the number after each action is an immediate reward. For example, $A, a_1, -3, A$ means that the agent took action a_1 from state A, received an immediate reward -3 and ended up in state A.

1. [9 Pts] Using a learning rate of $\alpha = 0.1$, and assuming initial state values of 0, what updates to $V(A)$ does *on-line* TD(0) method make after the first episode?

2. [10 Pts] Draw a diagram of an MDP and policy π that can best fit these two episodes (i.e., the maximum likelihood Markov model). Write down the transition probability $P^a_{s,s'}$, the reward function R^a_s of the MDP you draw, and your estimated policy $\pi(a|s)$.

3. [6 Pts] Based on your results in the previous question, solve Bellman equation to find out the state-value function of $V_\pi(A)$ and $V_\pi(B)$. (Assume $V_\pi(C) = 0$ as state C is the terminal state.)

4. [5 Pts] What value function would batch TD(0) find, i.e., if TD(0) was applied repeatedly to these two episodes?