Midterm Exam Sample

ELEN E6885: Introduction to Reinforcement Learning

Problem 1 (2 Points each)

True or False. No explanation is needed.

1. In multi-arm bandit problem, the incremental implementation (rather than taking average) is computationally efficient in terms of constant memory and constant per-time-step computation. Solution: True

2. For a finite MDP, the policy iteration process must converge to an optimal policy and optimal value function in a finite number of iterations. Solution: True

3. Consider the value functions v_k and v_{k+1} from two iterations of value iteration. Let π_k and π_{k+1} be the policies that are greedy with respect to v_k and v_{k+1}, respectively. It is always true that $\pi_{k+1} \geq \pi_k$, i.e., $v_{\pi_{k+1}}(s) \geq v_{\pi_k}(s)$ for any state s. Solution: False

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

5. 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. Solution: True

Problem 2 (5 Points each)

Short-answer questions.

1. Consider a k-armed bandit problem with $k = 4$ actions, denoted $1, 2, 3$ and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$ for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = 1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = 2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred? A_2 and A_5 were definitely exploratory. Any of the others could have been exploratory.

   Rubrics: 1 point to each correct statement on A_i.

2. Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions? The main thing about the MDP is Markov property. There are tasks where this does not hold. For instance, in Poker, the previous states will determine what is in the deck and what is not. This does not obey Markov property.

   Rubrics: 2 point for the first question. 3 points for the example.

3. What is the definition of one-step TD error δ_t using the state-value function? What is $\mathsf{E}[\delta_t|S_t = s]$ if δ_t uses the true state-value function v_π?

   Solution: One-step TD error is defined as $\delta_t = R_{t+1} + \gamma v(S_{t+1}) - v(S_t)$. If the true state-value function is used in δ_t, we have

$$\begin{aligned}\mathsf{E}[\delta_t|S_t = s] &= \mathsf{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) - v_\pi(S_t)|S_t = s\right] \\ &\overset{(a)}{=} v_\pi(s) - v_\pi(s) \\ &= 0,\end{aligned}$$

   where (a) is due to Bellman expectation equation.

   Rubrics: 2 points to give the correct one-step TD error formula; 1 more point to correctly write down $\mathsf{E}[\delta_t|S_t = s]$ using the true state-value function; 1 more point for answering $\mathsf{E}[\delta_t|S_t = s] = 0$.

Problem 3 (**10 Points**) Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1 - p$. Let the reward be $+1$ on all transitions, and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps, with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state? first-visit: $v_s = 10$; all-visit: $v_s = \frac{1}{10} * (1 + 2 + 3 + ... + 9 + 10) = 5.5$.

Rubrics: 5 points for each correct calculation.

Problem 4 (**10 Points**) Consider adding a constant c to all the rewards in an episodic task with discounted factor $\gamma \in (0, 1)$, such as maze running. Would this have any effect on the optimal policy? If yes, please give an example. If not, please give the proof.

Solution: Let terminal time be T. Then the new return at time t is given below $G_t \leftarrow G_t + c * \frac{1-\gamma^T}{1-\gamma}$. Suppose that we have an episodic task with one state S and two actions A_0, A_1. A_0 takes agent to terminal state with reward 1, while A_1 takes agent back to S with reward 0. In this case, the agent should terminate to maximize reward. However, if we add 1 to each reward then the return for doing A_1 forever is $\frac{1}{1-\gamma}$. Let $\gamma < \frac{1}{2}$, the return is bigger than 2. Therefore, the optimal policy is changed to take A_1 forever.

Rubrics: 2 points to "yes"; 3 points for the calculation of $G_t \leftarrow G_t + c * \frac{1-\gamma^T}{1-\gamma}$.; 5 points for the example.