Sample Midterm Exam

ELEN E6885: Introduction to Reinforcement Learning

Problem 1 (**20 Points, 4 Points each**)

Please indicate *True* or *False* in the following statements (no explanation needed):

1. The n-armed bandit problem in our lecture can be viewed as one-state MDP. True
2. Assume MRP model is known, then the bellman equation for MRP can be solved in a closed matrix form. However, direct solution is only possible for small MRPs. True
3. In the case each return is an independent, identically distributed estimate of the value function $v(s)$ with finite variance, the convergence of first-visit Monte Carlo method to the true value $v(s)$ as the number of visits of each s goes to infinity is guaranteed by following the law of large numbers. True
4. In general, Monte Carlo method is better on the existing data (e.g. rewards collected from the episodes) while TD methods produce lower error on future data. True
5. Expected Sarsa has higher variance in its updates than Sarsa. False

Problem 2 (**20 Points, 4 Points each**)

Answer the following questions:

1. In model-based reinforcement learning, what cases would you prefer to use approximated dynamic programming? Approximate dynamic programming is often used for problems with large or continuous state and action spaces. In these problems, dynamic programming may be intractable so an approximation to the optimal policy is often good enough.
2. In model-free Monte Carlo method, why is it useful to estimate Q value rather than V value for policy improvement? Policy improvement is done by using Q values. If only V values are known, we have to use the MDP model to compute Q values.
3. In reinforcement learning, what are the meanings of "on-policy" method and "off-policy" method? on-policy method attempts to evaluate or improve the policy that is used to make decisions. The off-policy method attempts to evaluate the policy by following another policy.

4. For Q-learning to converge (in terms of Q value) we need to correctly manage the exploration vs. exploitation tradeoff. What property needs to be hold for the exploration strategy? In the limit, every action needs to be tried sufficiently often in every possible state. Also, the exploration needs to be vanishing over time, e.g., decreasing ϵ to zero in ϵ-greedy.

5. Which of the following reinforcement learning algorithms estimate some quantity based on experience, from which we are NOT guaranteed to obtain the *optimal policy*? Choose one or more that apply and explanation is needed.
a) model-free Monte Carlo; b) Q-learning; c) TD(0) learning. Answer: a, c. Q-learning estimates the Q-value of the optimal policy, while the others estimate the Q-value of a fixed policy.

Problem 3 (**Small Grid World, 30 Points**)

Consider the small grid world MDP as shown in Fig. 1. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1). There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The underlying transition function is such that the intended agent movement (up, down, left, right) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. For example, if the agent at (1,2) chooses to move up to (2,2), it can reach (2,2) only with probability 0.8. The agent may hit (1,1) or terminal state -5 with equal probability 0.1. If a collision with a wall happens, the agent stays in the same state.

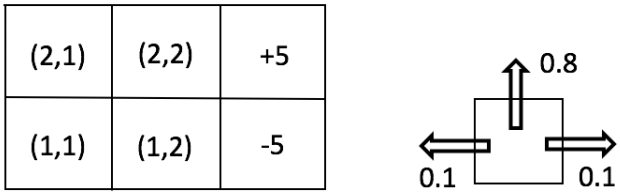

Figure 1: Grid world MDP and transition function

1. [4 Pts] Write down the optimal policy for this grid. (For example, $\pi(1,1) = down, \pi(1,2) = up, \pi(2,1) = left, \pi(2,2) = right$, this is definitely NOT an optimal policy.)

2. [8 Pts] Suppose the agent knows the transition probabilities. Give the first two rounds of (synchronous) *value iteration* updates for state (1,2) and (2,1), with a discount of 0.9. (Assume V_0 is 0 everywhere and compute V_i for times $i = 1, 2$. Also, we assume the values of terminal states $V(1,3) = V(2,3) = 0$ for all iterations).

3. [8 Pts] The agent starts with the policy that always chooses to go right, and executes the following three trials: 1) (1,1)-(1,2)-(1,3), 2) (1,1)-(1,2)-(2,2)-(2,3), and 3) (1,1)-(2,1)-(2,2)-(2,3). What are the Monte Carlo (direct utility) estimates for states (1,1) and (2,2), given these traces?

4. [10 Pts] Using a learning rate of $\alpha = 0.1$, a discount of $\gamma = 0.9$, and assuming initial V values of 0, what updates does the TD(0)-learning agent make after trials 1 and 2, above?

Problem 4 (Radom Walk, 30 Points)

See Figure 2. All episodes start in the center state, C, and proceed either left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or the extreme right. When an episode terminates on the right a reward of +1 occurs; all other rewards are zero. For example, a typical walk might consist of the following state-and reward sequence: $C, 0, B, 0, C, 0, D, 0, E, 1$. Assume this task is undiscounted ($\gamma = 1$) and episodic, then the true value of each state is essentially the probability of terminating on the right if starting from that state.

1. [3 Pts] What is the state value $V(C)$?

2. [12 Pts] Then what are the state values $V(A), V(B), V(D), V(E)$? Detailed computation is needed.

3. [15 Pts] Suppose we initialize state value $V(s) = 0.5$ for all state s. From Figure 3 (the learning process), it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?

Figure 2: Random Walk

1. $V(C) = \frac{1}{2}$. Hint: the probabilities of terminating at left and right are equal. We earn reward 1 by terminating at right. Thus, $V(C) = \frac{1}{2}$.

2. We know $V(C) = \frac{1}{2}$. Then according to the Bellman expectation equation,

$$\begin{aligned} V(A) &= 0.5V(B) \\ V(B) &= 0.5V(C) + 0.5V(A) \\ V(D) &= 0.5V(C) + 0.5V(E) \\ V(E) &= 0.5(1+0) + 0.5V(D) \end{aligned} \tag{1}$$

Then, $V(A) = \frac{1}{6}, V(B) = \frac{2}{6}, V(D) = \frac{4}{6}, V(E) = \frac{5}{6}$.

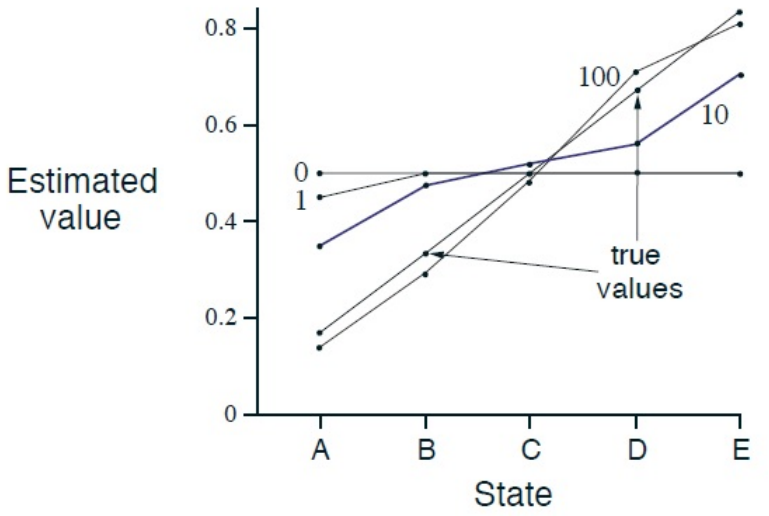

Figure 3: Values learned by TD(0) with a constant step-size $\alpha = 0.1$ after various numbers of episodes

3. In the first episode the random walk took us off the left end of the domain and we decreased the value function of state A by 0.05 (i.e. $V(A) = 0.45$). It directly follows from TD(0) updating formula $V(s_t) \leftarrow V(s_t) + \alpha(R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$ where $\gamma = 1$.