Reinforcement Learning

End Semester Exam

17/12/2022

Sanjit K. Kaul (with inputs from Samaksh Gupta)

Instructions: You have two hours to work on the questions. Answers with no supporting steps will receive no credit. No resources, other than a pen/pencil, are allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 10 **marks** Show the steps that help obtain the optimal policy from the optimal state-value function.

Question 2. 20 **marks** Derive $E[G_{t+1}|S_t]$ in terms of the rewards at $t+1$ and $t+2$, state at $t+2$, the value function $v_\pi(s)$, the MDP $p(s', r|s, a)$ and the policy PMF $\pi(a|s)$. Assume a discount factor γ.

Let $P_t(s)$ be the PMF over the set of states at time t. Use the expression obtained above and $P_t(s)$ to calculate $E[G_{t+1}]$. Simplify the expression under the assumption that the policy is a deterministic policy $\pi(s)$.

Question 3. 20 **marks** Suppose you are given the state-value function $v_\pi(s)$. Use $v_\pi(s)$ and the MDP to calculate a policy π' that is as good as the policy π. You must provide a convincing argument as to why your calculations will result in a policy π' that is at least as good as π.

Argue that if $v_\pi(s) = v_{\pi'}(s)$, for all states s, then π is an optimal policy. You want to use the fact that the optimal value function satisfies the Bellman optimality equations.

Question 4. 20 **marks** Suppose the state of an environment is an n-dimensional vector. Further assume that each element of the vector can take a value of 0 or 1. What is the dimensionality of the state-value vector v_π?

Provide an orthonormal basis that can be used to exactly express the state-value vector as a linear combination of one or more (column) vectors in the basis.

Provide a linear approximation architecture that approximates the state-value vector using d vectors in the basis. Write down the mean-squared error for your architecture. Derive the approximation that minimizes the error.

In practice you can't calculate the linear approximation as above, since you don't know $v_\pi(s)$. Assume a suitable data-driven sample approximation for $v_\pi(s)$ and write down the corresponding sample estimate of the squared error.

Use the sample estimate to calculate the gradient descent step that you will iterate over till the iterations converge to an approximation.

Question 5. 5 **marks** Rewrite the following expression as an expected value over the pair (S, A) of random variables.

$$\sum_s \sum_a \mu(s) \nabla \pi_\theta(a|s) q_{\pi_\theta(s,a)}.$$

Question 6. 10 **marks** Show that if the critic approximation $f_w(s,a)$ satisfies the conditions

$$\sum_s \rho_\pi(s) \sum_a \pi(a|s,\theta)(Q_\pi(s,a) - f_w(s,a))\nabla_w f_w(s,a) = 0,$$
$$\nabla_w f_w(s,a) = \nabla_\theta \pi(a|s,\theta)\frac{1}{\pi(a|s,\theta)},$$

then

$$\sum_s \sum_a \rho_\pi(s)\nabla_\theta \pi(a|s,\theta)Q_\pi(s,a) = \sum_s \sum_a \rho_\pi(s)\nabla_\theta \pi(a|s,\theta)f_w(s,a).$$

What $f_w(s,a)$ satisfies $\nabla_w f_w(s,a) = \nabla_\theta \pi(a|s,\theta)\frac{1}{\pi(a|s,\theta)}$? Explain your answer.

Question 7. 15 **marks** Consider the expected discounted sum reward

$$\eta(\pi) = E_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]. \tag{1}$$

Show that for an alternate policy π',

$$\eta(\pi') = \eta(\pi) + E_{s_0,a_0,\cdots\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right],$$

where $A_\pi(s_t,a_t) = Q_\pi(s_t,a_t) - v_\pi(s_t)$, $s_0 \sim \rho_0(s_0)$.

Question 8. 30 **marks; For extra credit.** You have three possible states M1, M2, and M3. You have two possible choices: Attempt A1 and Not Attempt A2. When you are in state M1 and choose Attempt, you stay in M1 with a probability of 0.9 and transition to M2 with a probability of 0.1. Staying in M1 gets a reward of 4, and transitioning into M2 gives you a reward of 2.

If you choose to Not Attempt in state M1, you transition to M3 with probability 0.5 and get a reward of -2. With the remaining probability, you stay in M1 and get a reward of 0.

Starting in M2 and choosing to Attempt can lead you to M1 with a probability of 0.5 and a reward of 2. With probability 0.4 it can lead you to M3, and that will earn a reward of -2. With the remaining probability, you will stay in M2 and get a reward of 0.

If you choose to Not Attempt, you can transition to state M3 with a probability of 0.6, and you'll get a reward of -2. With probability 0.4, you will remain in the M2 state and get a reward of 0.

Starting in M3 and choosing to Attempt can result in M1 with probability 0.4 and a reward of 4. With the remaining probability, you will stay in M3 and get 0 reward.

If you choose to Not Attempt, you can transition to the M1 state with a probability of 0.6, and you'll get a reward of 2. With probability 0.3, you will remain in the M3 state and get a reward of 0. With the remaining probability, you can transition to M2 and get a reward of 1.

Answer the following questions.

(a) (5 marks) Draw the MDP.

(b) (15 marks) Begin with a policy that chooses actions with equal probability in every state. Carry out policy iteration. Show two iterations. State your obtained policy and its value function ($v_\pi(s)$ for all s) on completion of the iterations. Assume a discount factor of 0.5. [Hint: Using value iteration or evaluating policy via iteration will result in zero credit.]