Reinforcement Learning

Final Exam

11/12/2021

Sanjit K. Kaul

Instructions: You have two hours to work on the questions and an additional ten minutes to upload a single PDF containing your work. Please ensure that you give uploading sufficient time. Delayed uploads will not be graded. **Final answers with no supporting steps will receive no credit**. Exam is open book and class notes. No other resources, human or otherwise are allowed. Violation of the policy will be treated as use of unfair means and plagiarism. All academic penalties will apply.

Question 1. 20 **marks** Our environment has one state 0 and a terminal state Z. In state 0 one can take the two actions of -1 and 1. On taking either action the environment transitions to 0 with probability 0.2 and the agent gets a reward of 0. With probability 0.8, the agent gets a reward 1 and the episode terminates. You want to estimate the value of state 0 when using the policy π that picks action 1 with probability 1. However, you only have access to the policy μ that picks either action with probability 0.5. You are allowed to run one episode. Derive the expected value and variance of your estimate of $v_\pi(0)$. Assume a discount factor of 1.

Question 2. 10 **marks** Our environment has two states s_0 and s_1 and two terminal states. Five episodes obtained by using a policy π are listed next.

$$s_0\ 0.5\ s_1\ 0,$$
$$s_0\ 0.2,$$
$$s_0\ 0.3,$$
$$s_1\ 0.6,$$
$$s_1\ 0.4.$$

The episodes above are summarised as $S_t, R_{t+1}, S_{t+1}, R_{t+2}, \ldots$. You are interested in estimating the values of the states. You can use either TD(0) or first visit MC. Derive the estimates both will converge to. Explain your steps.

Question 3. 20 **marks** Our environment has two states A and B and actions -1 and 1. Z is the terminal state. Consider the following episode, where in an episode is summarised as $(S_t, A_t, R_{t+1}), (S_{t+1}, A_{t+1}, R_{t+2}), \ldots$. The episode is

$$(A, 1, 5), (B, -1, 5), (B, 1, 3), (A, -1, -5), (B, 1, 5), Z.$$

Use the episode to derive the action value estimates one would obtain using SARSA and Q-learning.

Question 4. 20 **marks** Consider the following experiences, wherein an experience is in the form $S_t, A_t, R_{t+1}, S_{t+1}$.

A and B are states and Z is the terminal state.

$$A, -1, 3, A,$$
$$A, -1, 5, B,$$
$$A, 1, 2, A,$$
$$B, -1, 4, Z,$$
$$B, 1, 3, A.$$

Assume initial action-values of $Q(A, 1) = 2$, $Q(A, -1) = 3$, $Q(B, 1) = 2$, $Q(B, -1) = 3$. Use these to associate with each experience a TD error. Calculate the probability with which an experience would be picked when using (a) greedy TD error experience replay, (b) uniform replay, and (c) rank based prioritization. Show all your steps.

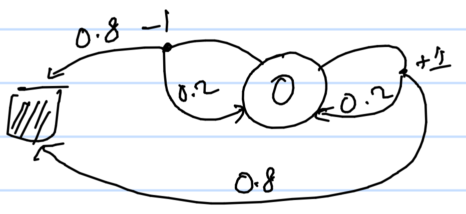
Question 5. 10 **marks** We have states $1, 2, \ldots, n$. Assume a linear approximation architecture. State i is mapped to a n-dimensional feature vector with the i^{th} element set to 1 and the other elements set to 0. Suppose we observe an episode with the sequence $1, 2, \ldots, n$ of states. Derive the eligibility trace vector z_n for the episode. Assume TD(0.3) and a discount factor of 0.8.

Question 6. 20 **marks** Consider $r(\theta)$, where r is the average reward and θ parameterizes the policy π. Write down $r(\theta)$ in terms of the parameterized policy and the true action value function $q_\pi(s, a)$. Further write down the gradient of $r(\theta)$ with respect to θ. Assume the steady state distribution is given by $\mu_\pi(s)$.

Assume that we approximate the value function (as we would in an actor-critic method) using a linear approximation architecture with the feature vector for (s, a) given by $\nabla_\theta \log(\pi(a|s, \theta))$ and weight vector w. We want to minimize the average squared error in estimating $r(\theta)$ when using the linear approximation in place of the true action value function. Write down the expression for the error and the necessary first order condition that the error minimizing w must satisfy.

Derive the gradient of $r(\theta)$ in terms of the optimal linear approximation and show that it is the same as the gradient that you defined earlier using the true action-value function.

Question 1. 20 marks Our environment has one state 0 and a terminal state Z. In state 0 one can take the two actions of -1 and 1. On taking either action the environment transitions to 0 with probability 0.2 and the agent gets a reward of 0. With probability 0.8, the agent gets a reward 1 and the episode terminates. You want to estimate the value of state 0 when using the policy π that picks action 1 with probability 1. However, you only have access to the policy μ that picks either action with probability 0.5. You are allowed to run one episode. Derive the expected value and variance of your estimate of $v_\pi(0)$. Assume a discount factor of 1.



Your target policy is π, wherein $\pi(1|0) = 1$.

The behavior policy is μ, wherein $\mu(1|0) = \mu(-1|0) = 0.5$.

Note that any episode has a return of 1.

However, when using the behavior policy, an episode will contain a random T steps.

The mean of our estimate is $E_\mu\left[\rho_{0:T-1} G_0 \mid S_0 = 0\right]$

$$= E_\mu\left[\rho_{0:T-1} G_0\right] \quad \left\{ S_0 = 0 \text{ is implicit} \right.$$

$$= \sum_{x=1}^{\infty} E_\mu\left[\rho_{0:T-1} G_0 \mid T = x\right] P\left[T = x\right]$$

$$= \sum_{x=1}^{\infty} E_\mu\left[\rho_{0:T-1} \mid T = x\right] P\left[T = x\right]$$

$$P[T=1] = 2(0.5)(0.8) = 0.8$$

$$P[T=2] = (1-0.8)(0.8) = (0.2)(0.8)$$

$$P(T=3) = (0.2)^2(0.8)$$

$$\vdots$$

$$P(T=x) = (0.2)^{x-1}(0.8)$$

3

$$\rho_{0:x-1} = \frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \frac{\pi(a_1|s_1)}{\mu(a_1|s_1)} \cdots \frac{\pi(a_{x-1}|s_{x-1})}{\mu(a_{x-1}|s_{x-1})}$$

$$= \begin{cases} 0 & \text{otherwise.} \\ \dfrac{1}{(0.5)^x} & a_0 = a_1 = \cdots = a_{x-1} = 1. \end{cases}$$

3

$$E_\mu\left[\rho_{0:x-1}\right] = \frac{1}{(0.5)^x}(0.5)^x = 1.$$

Here we use the fact that $P\left[a_0 = a_1 = \cdots = a_{x-1} = 1\right] = (0.5)^x$ when using μ.

$$\therefore E_\mu\left[\rho_{0:T-1} G_0\right] = 1(0.8) + 1(0.2)(0.8) + \cdots$$

$$= (0.8)(1 + 0.2 + (0.2)^2 + \cdots)$$

$$= (0.8)\left(\frac{1}{1-0.2}\right) = 1.$$

3

The variance of the estimate is

$$Var\left[\rho_{0:T-1} G_0\right] = E_\mu\left[(\rho_{0:T-1} G_0)^2\right] - \left(E_\mu\left[\rho_{0:T-1} G_0\right]\right)^2$$

$$E_\mu\left[(\rho_{0:T-1} G_0)^2\right] = (1)^2(0.8) + (1)^2(0.2)(0.8) + \cdots$$

$$= 1.$$

$$\therefore Var[\quad] = 0.$$

3

This makes sense as the behavior policy can only result in a return of 1 at the end of any episode.

Question 2. 10 **marks** Our environment has two states s_0 and s_1 and two terminal states. Five episodes obtained by using a policy π are listed next.

$$s_0\ 0.5\ s_1\ 0,$$
$$s_0\ 0.2,$$
$$s_0\ 0.3,$$
$$s_1\ 0.6,$$
$$s_1\ 0.4.$$

The episodes above are summarised as $S_t, R_{t+1}, S_{t+1}, R_{t+2}, \ldots$. You are interested in estimating the values of the states. You can use either TD(0) or first visit MC. Derive the estimates both will converge to. Explain your steps.

Recall that TD(0) converges to certainty equivalent estimates.
Given the episodes above, a transition from s_1 results in the episode being terminated. The returns starting in s_1 are 0, 0.6, 0.4.

$$\therefore\ V_\pi^{TD(0)}(s_1) \longrightarrow \frac{0+0.6+0.4}{3} = 1/3$$

7

As regards s_0:
We have s_0 transitioning to s_1 with probability 1/3 (as determined from the data).

We have
$$V_\pi^{TD(0)}(s_0) = \frac{1}{3}\left(0.5 + V_\pi^{TD(0)}(s_1)\right) + \frac{1}{3}(0.2) + \frac{1}{3}(0.3)$$
$$= \frac{1}{3}\left(0.5 + \frac{1}{3}\right) + \frac{1}{3}\left(\frac{1}{2}\right)$$
$$= \frac{1}{3} + \frac{1}{9} = \frac{4}{9}.$$

For MC

$$V_\pi^{MC}(s_0) = \frac{0.5+0.2+0.3}{3} = 1/3.$$

3

$$V_\pi^{MC}(s_1) = 1/3.$$

Question 3. 20 marks Our environment has two states A and B and actions -1 and 1. Z is the terminal state. Consider the following episode, where in an episode is summarised as $(S_t, A_t, R_{t+1}), (S_{t+1}, A_{t+1}, R_{t+2}), \ldots$. The episode is

$$(A, 1, 5), (B, -1, 5), (B, 1, 3), (A, -1, -5), (B, 1, 5), Z.$$

Use the episode to derive the action value estimates one would obtain using SARSA and Q-learning.

$\alpha = 0.2$ [α must be picked from $(0,1)$]
$\gamma = 1$

10

Initial Q-values are all set to zero.

$Q(A,-1)=0$
$Q(A,1)=0$
$Q(B,-1)=0$
$Q(B,1)=0$

Q-learning:

$Q(A,1) = Q(A,1) + \alpha(5 + \gamma Q(B, \underset{a'}{\text{argmax}}\, Q(B,a')))$
$= 0 + (0.2)(5+0)$
$= 1.0.$

We have $Q(A,-1)=0$, $Q(A,1)=1$, $Q(B,-1)=0$, $Q(B,1)=0$.

$Q(B,-1) = Q(B,-1) + \alpha(5 + \gamma Q(B, \underset{a'}{\text{argmax}}\, Q(B,a')))$
$= 0 + 0.2(5+0) = 1.0.$

We have $Q(A,-1)=0$, $Q(A,1)=1$, $Q(B,-1)=1$, $Q(B,1)=0$.

$Q(B,1) = Q(B,1) + \alpha(3 + \gamma \max_{a'} Q(A,a'))$
$= 0 + (0.2)(3+1)$
$= (0.2)(4) = 0.8.$

We have $Q(A,-1)=0$, $Q(A,1)=1$, $Q(B,-1)=1$, $Q(B,1)=0.8$

$Q(A,-1) = Q(A,-1) + \alpha(-5 + \gamma \max_{a'} Q(B,a'))$
$= 0 + (0.2)(-5+1)$
$= (0.2)(-4) = -0.8.$

We have:
$Q(A,-1) = -0.8$, $Q(A,1)=1$, $Q(B,-1)=1$, $Q(B,1)=0.8$

$Q(B,1) = Q(B,1) + \alpha(5 + \gamma(0))$
$= (0.8) + (0.2)(5) = 1.8.$

We have: $Q(A,-1) = -0.8$
$Q(A,1) = 1$
$Q(B,-1) = 1$
$Q(B,1) = 0.8$

SARSA

10

For reference: $(A, 1, 5), (B, -1, 5), (B, 1, 3), (A, -1, -5), (B, 1, 5), Z.$

Again $Q(A,-1)=0$
$Q(A,1)=0$
$Q(B,-1)=0$
$Q(B,1)=0$

$Q(A,1) = Q(A,1) + \alpha(5 + \gamma Q(B,-1))$
$= 0 + (0.2)(5) = 1.0.$

We have:
$Q(A,-1)=0$, $Q(A,1)=1.0$, $Q(B,-1)=0$, $Q(B,1)=0$

$Q(B,-1) = Q(B,-1) + \alpha(5 + \gamma Q(B,1))$
$= 0 + (0.2)(5+0)$
$= 1.$

We have:
$Q(A,-1)=0$, $Q(A,1)=1.0$, $Q(B,-1)=1$, $Q(B,1)=0$

$Q(B,1) = Q(B,1) + \alpha(3 + \gamma Q(A,-1))$
$= (0.2)(3+0) = 0.6.$

We have:
$Q(A,-1)=0$, $Q(A,1)=1.0$, $Q(B,-1)=0$, $Q(B,1)=0.6$

$Q(A,-1) = Q(A,-1) + \alpha(-5 + \gamma Q(B,1))$
$= (0.2)(-5+0.6)$
$= -1 + 0.12 = -0.88$

We have:
$Q(A,-1)=-0.88$, $Q(A,1)=1.0$, $Q(B,-1)=0$, $Q(B,1)=0.6$

$Q(B,1) = Q(B,1) + \alpha(5+0)$
$= 0.6 + (0.2)(5) = 1.6.$

We have:
$Q(A,-1)=-0.88$, $Q(A,1)=1.0$, $Q(B,-1)=0$, $Q(B,1)=1.6$

Question 4. 20 **marks** Consider the following experiences, wherein an experience is in the form $S_t, A_t, R_{t+1}, S_{t+1}$.

A and B are states and Z is the terminal state.

$$A, -1, 3, A,$$
$$A, -1, 5, B,$$
$$A, 1, 2, A,$$
$$B, -1, 4, Z,$$
$$B, 1, 3, A.$$

Assume initial action-values of $Q(A, 1) = 2$, $Q(A, -1) = 3$, $Q(B, 1) = 2$, $Q(B, -1) = 3$. Use these to associate with each experience a TD error. Calculate the probability with which an experience would be picked when using (a) greedy TD error experience replay, (b) uniform replay, and (c) rank based prioritization. Show all your steps.

Consider the first experience

$\boxed{A, -1, 3, A}$

TD-error $\delta_1 = R_1 + \gamma Q(A, \arg\max_a Q(A,a)) - Q(A,-1)$

$= 3 + Q(A,-1) - Q(A,-1) = 3.$

$\boxed{A, -1, 5, B}$

$\delta_2 = R_2 + \gamma Q(B, \arg\max_a Q(B,a)) - Q(A,-1)$

$= 5 + Q(B,-1) - Q(A,-1)$

$= 5 + 3 - 3 = 5.$

5

$\boxed{A, 1, 2, A}$

$\delta_3 = 2 + Q(A,-1) - Q(A,1)$

$= 2 + 3 - 2 = 3.$

$\boxed{B, -1, 4, Z}$

$\delta_4 = 4 + 0 - Q(B,-1)$

$= 4 - 3$

$= 1.$

$\boxed{B, 1, 3, A}$

$\delta_5 = 3 + Q(A,-1) - Q(B,1)$

$= 3 + 3 - 2 = 4.$

7.5

To summarize:

	Experience	TD-error	Rank	1/Rank	P[Rank-based picking of transition i] ($\alpha = 0.2$)
(1)	A, -1, 3, A	$\delta_1 = 3$	3	1/3	$(1/3)^\alpha / \left(\frac{1}{3}^\alpha + 1^\alpha + \frac{1}{3}^\alpha + \frac{1}{4}^\alpha + \frac{1}{2}^\alpha\right) = 0.18$
(2)	A, -1, 5, B	$\delta_2 = 5$	1	1	0.23
(3)	A, 1, 2, A	$\delta_3 = 3$	3	1/3	0.18
(4)	B, -1, 4, Z	$\delta_4 = 1$	4	1/4	0.18
(5)	B, 1, 3, A	$\delta_5 = 4$	2	1/2	0.20

Greedy:
Pick the experience with the largest error w.p. 1. So we will pick experience (5).

Uniform: P[picking any experience] = 1/5.

7.5

Rank-based:

$$P[\text{Sampling transition } i] = \frac{p_i^\alpha}{\sum_{k=1}^{5} p_k^\alpha}$$

$$p_i = \frac{1}{\text{rank}(i)}$$

The ranks and the corresponding probabilities are listed in the table above.

Question 5. 10 **marks** We have states $1, 2, \ldots, n$. Assume a linear approximation architecture. State i is mapped to a n-dimensional feature vector with the i^{th} element set to 1 and the other elements set to 0. Suppose we observe an episode with the sequence $1, 2, \ldots, n$ of states. Derive the eligibility trace vector z_n for the episode. Assume TD(0.3) and a discount factor of 0.8.

Initialize $z_0 = \bar{0}$, where $\bar{0}$ is the n-dimensional zero vector.

We have $s_1 = 1, s_2 = 2, \ldots, s_n = n$

$$z_1 = \gamma\lambda z_0 + \nabla \hat{v}(s_1, \bar{w}_1)$$
$$= \nabla \hat{v}(s_1, \bar{w}_1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1}$$

2

$$z_2 = \gamma\lambda z_1 + \nabla \hat{v}(s_2, \bar{w}_2) = \gamma\lambda \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$z_3 = \gamma\lambda z_2 + \nabla \hat{v}(s_3, \bar{w}_3) = (\gamma\lambda)^2 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + (\gamma\lambda) \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\vdots$

$$z_n = \begin{bmatrix} (\gamma\lambda)^{n-1} \\ (\gamma\lambda)^{n-2} \\ \vdots \\ (\gamma\lambda) \end{bmatrix} = \begin{bmatrix} (0.24)^{n-1} \\ (0.24)^{n-2} \\ \vdots \\ (0.24) \end{bmatrix}$$

8

Question 6. **20 marks** Consider $r(\theta)$, where r is the average reward and θ parameterizes the policy π. Write down $r(\theta)$ in terms of the parameterized policy and the true action value function $q_\pi(s,a)$. Further write down the gradient of $r(\theta)$ with respect to θ. Assume the steady state distribution is given by $\mu_\pi(s)$.

Assume that we approximate the value function (as we would in an actor-critic method) using a linear approximation architecture with the feature vector for (s,a) given by $\nabla_\theta \log(\pi(a|s,\theta))$ and weight vector w. We want to minimize the average squared error in estimating $r(\theta)$ when using the linear approximation in place of the true action value function. Write down the expression for the error and the necessary first order condition that the error minimizing w must satisfy.

Derive the gradient of $r(\theta)$ in terms of the optimal linear approximation and show that it is the same as the gradient that you defined earlier using the true action-value function.

$$r(\theta) = \sum_s \mu_\pi(s) \sum_a \pi(a|s,\theta)\, q_\pi(s,a)$$

The gradient $\nabla r(\theta)$ is given by the policy gradient theorem for the average reward case. We have:

$$\nabla r(\theta) = \sum_s \mu(s) \sum_a q_\pi(s,a)\, \nabla \pi(a|s,\theta)$$

5

Let the linear approximation be

$$f_w(s,a) = \nabla_\theta \log(\pi(a|s,\theta))^T w$$

The approximation of $r(\theta)$ is

$$\hat{r}(\theta) = \sum_s \mu(s) \sum_a \pi(a|s,\theta)\, f_w(s,a)$$

The average squared error in approximation is:

$$\sum_s \mu(s) \sum_a \pi(a|s,\theta) \left[(q_\pi(s,a) - f_w(s,a))^2 \right]$$

The necessary first order conditions that the w that minimizes the above must satisfy are

5

$$\sum_s \mu(s) \sum_a \pi(a|s,\theta)\,(-2)\,(q_\pi(s,a) - f_w(s,a))\, \nabla_w f_w(s,a) = 0$$

We have

$$\sum_s \mu(s) \sum_a \pi(a|s,\theta)\, q_\pi(s,a)\, \nabla_w f_w(s,a) = \sum_s \mu(s) \sum_a \pi(a|s,\theta)\, f_w(s,a)\, \nabla_w f_w(s,a)$$

Note that given the linear approximation architecture that f_w is, we have

$$\nabla_w f_w(s,a) = \nabla_\theta \log \pi(a|s,\theta) = \frac{\nabla_\theta \pi(a|s,\theta)}{\pi(a|s,\theta)}.$$

5

Substituting $\nabla_w f_w(s,a)$ in the equation above corresponding to first order conditions, we get

5

$$\sum_s \mu(s) \sum_a q_\pi(s,a)\, \nabla_\theta \pi(a|s,\theta) = \sum_s \mu(s) \sum_a f_w(s,a)\, \nabla_\theta \pi(a|s,\theta)$$

Which is to say that the $\nabla r(\theta)$ calculated using $q_\pi(s,a)$ is the same as that obtained on replacing $q_\pi(s,a)$ by the approximation architecture $f_w(s,a)$. //