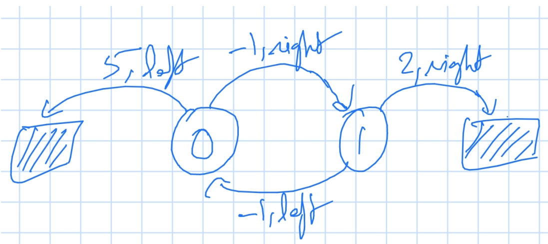
Reinforcement Learning

Mid Semester Exam

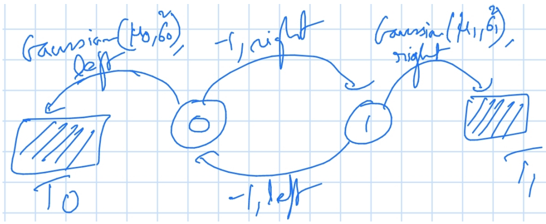
23/09/2023

Sanjit K. Kaul

Instructions: You have sixty minutes to work on the questions. Answers with no supporting steps will receive no credit. Any resources, other than a pen/pencil, are **not** allowed. In case you believe that required information is unavailable, make a suitable assumption.



(a) MDP



(b) MDP. We have two Gaussian random variables, each described by its mean and variance.

Fig. 1: MDPs you will use in the questions that follow. In any state, an agent may choose either left or right. The number on an arrows besides the action is the obtained reward.

Question 1. 20 **marks** Consider the MDP in Figure 1a. We have two policies π and μ. The policy PMF π is described by the probabilities $\pi(\text{left}|0) = 0.3$ and $\pi(\text{right}|1) = 0.8$. The policy μ is described by the probabilities $\pi(\text{left}|0) = 0.5$ and $\pi(\text{right}|1) = 0.5$. Does μ improve π or not? Support your claim appropriately. Assume $\gamma = 1$.

Question 2. 20 **marks** Derive an expression for the expected return when using policy π in terms of rewards, the MDP functions $p(s', r|s, a)$, and the policy PMF(s) $\pi(a|s)$. Use the derived expression to write the expected return when using policy μ in terms of the expected return when using π.

Question 3. 30 **marks** Consider the MDP in Figure 1a. Derive the optimal policy for $\gamma = 1$. *[Hint: Solve the Bellman Optimality Equations directly.]* Next consider a policy μ that chooses left or right in any state with equal probability. Below is an episode generated using μ. Each tuple has a state, action chosen in the state, and the resulting reward.

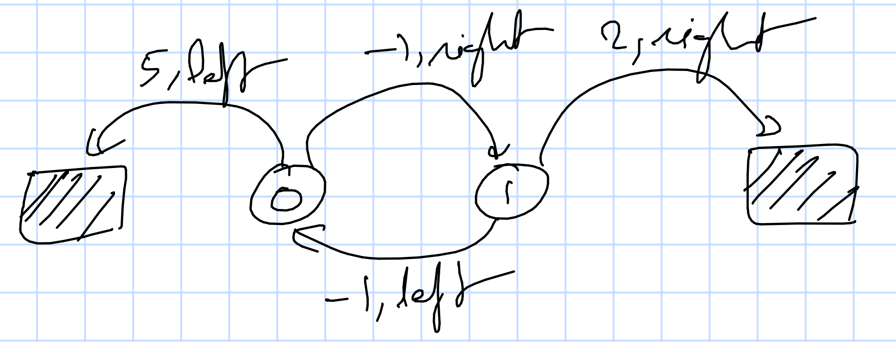
$$(0, \text{right}, -1),\ (1, \text{left}, -1),\ (0, \text{right}, -1),\ (1, \text{left}, -1),$$
$$(0, \text{right}, -1),\ (1, \text{left}, -1),\ (0, \text{right}, -1),\ (1, \text{right}, 2),\ \text{T}.$$

Estimate the value function for the policy μ using the episode. Apply every-visit Monte Carlo. Assume $\gamma = 1$ and use the sample mean. Now use the above episode generated using μ to estimate the value function of the optimal policy you derived above (*Hint: this is related to Question 2*).

Question 4. 30 **marks** Consider the MDP in Figure 1b. As always, we want the agent to take actions that maximize expected return. Also, we want to structure the rewards such that the actions that maximize the return have the agent end in terminal state T_1, starting from either 0 or 1. Provide conditions on $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$ that will satisfy our requirements. Assume that the discounting factor γ can take any value in $(0, 1]$.

Question 1. **20 marks** Consider the MDP in Figure 1a. We have two policies π and μ. The policy PMF π is described by the probabilities $\pi(\text{left}|0) = 0.3$ and $\pi(\text{right}|1) = 0.8$. The policy μ is described by the probabilities $\pi(\text{left}|0) = 0.5$ and $\pi(\text{right}|1) = 0.5$. Does μ improve π or not? Support your claim appropriately. Assume $\gamma = 1$.

We must evaluate the two policies. The MDP is:



$$v_\pi(0) = \pi(\text{left}|0)[5] + \pi(\text{right}|0)[-1 + v_\pi(1)]$$
$$= 1.5 + 0.7(-1 + v_\pi(1))$$
$$= 0.8 + 0.7\, v_\pi(1)$$

$$v_\pi(1) = \pi(\text{right}|1)(2) + \pi(\text{left}|1)[-1 + v_\pi(0)]$$
$$= (0.8)(2) + (0.2)(-1 + v_\pi(0))$$
$$= 1.6 - 0.2 + 0.2\, v_\pi(0)$$
$$= 1.4 + 0.2\, v_\pi(0)$$

$$\therefore v_\pi(1) = 1.4 + 0.2(0.8 + 0.7\, v_\pi(1))$$
$$= 1.4 + 0.16 + 0.14\, v_\pi(1)$$

$$0.86\, v_\pi(1) = 1.56$$
$$v_\pi(1) = \frac{1.56}{0.86} \quad \text{— } (1)$$

$$v_\pi(0) = 0.8 + 0.7\left(\frac{1.56}{0.86}\right) \quad (2)$$

$$v_\mu(0) = 0.5[5 + (-1) + v_\mu(1)]$$

$$v_\mu(1) = 0.5[2 + (-1) + v_\mu(0)]$$

$$v_\mu(0) = 2 + 0.5\, v_\mu(1)$$
$$v_\mu(1) = 0.5 + 0.5\, v_\mu(0)$$
$$= 0.5[1 + 2 + 0.5\, v_\mu(1)]$$
$$= 0.5[3 + 0.5\, v_\mu(1)]$$
$$= 1.5 + 0.25\, v_\mu(1)$$

$$0.75\, v_\mu(1) = 1.5$$
$$v_\mu(1) = 2$$

$$\therefore v_\mu(0) = 2 + 0.5(2)$$
$$= 3.$$

Observe that:
$$v_\pi(1) < v_\mu(1)$$
$$v_\pi(0) < v_\mu(0).$$

$\therefore \mu$ improves π.

Evaluation of μ 7.5
π 7.5
5 for the inference.

Question 2 15 + 5

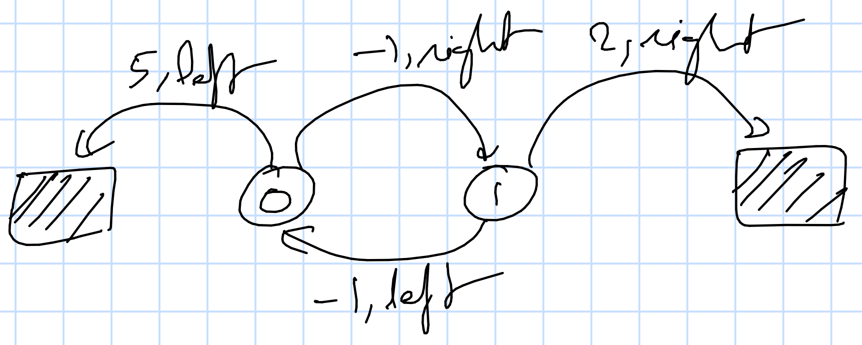
[We did this in class when deriving the importance sampling factor]

Essentially a freebie

Question 3. 30 **marks** Consider the MDP in Figure 1a. Derive the optimal policy for $\gamma = 1$. *[Hint: Solve the Bellman Optimality Equations directly.]* Next consider a policy μ that chooses left or right in any state with equal probability. Below is an episode generated using μ. Each tuple has a state, action chosen in the state, and the resulting reward.

$$(0, \text{right}, -1),\ (1, \text{left}, -1),\ (0, \text{right}, -1),\ (1, \text{left}, -1),$$
$$(0, \text{right}, -1),\ (1, \text{left}, -1),\ (0, \text{right}, -1),\ (1, \text{right}, 2),\ \text{T}.$$

Estimate the value function for the policy μ using the episode. Apply every-visit Monte Carlo. Assume $\gamma = 1$ and use the sample mean. Now use the above episode generated using μ to estimate the value function of the optimal policy you derived above (*Hint: this is related to Question 2*).



The optimality equations are:

$$V_*(0) = \max\{5, -1 + V_*(1)\}$$

$$V_*(1) = \max\{2, -1 + V_*(0)\}$$

Note that $V_*(1)$ has to satisfy $V_*(1) < 5$, given the MDP.

$\therefore\ V_*(0) = 5$

& $V_*(1) = \max\{2, -1+5\}$
$= 4.$

10

$\mu_*(0) = \text{left}$
$\mu_*(1) = \text{left}.$ / μ_* is the optimal policy

$$V_\mu(0) \simeq \frac{-5 + (-3) + (-1) + (1)}{4}$$

$$= \frac{-8}{4} = -2.$$

Each number in the numerator is a return corresponding to state 0.

10

$$V_\mu(1) \simeq \frac{-4 + (-2) + (0) + (2)}{4}$$

$$= \frac{-4}{4} = -1.$$

For state 0, the every-visit returns are $-5, -3, -1, 1$.

For state 1, the returns are $-4, -2, 0, 2$.

These returns were obtained using μ.

We would like to use the episode & the returns to estimate the values $V_*(0)$ and $V_*(1)$, which are the values of the states when the optimal policy μ_* is used.

10

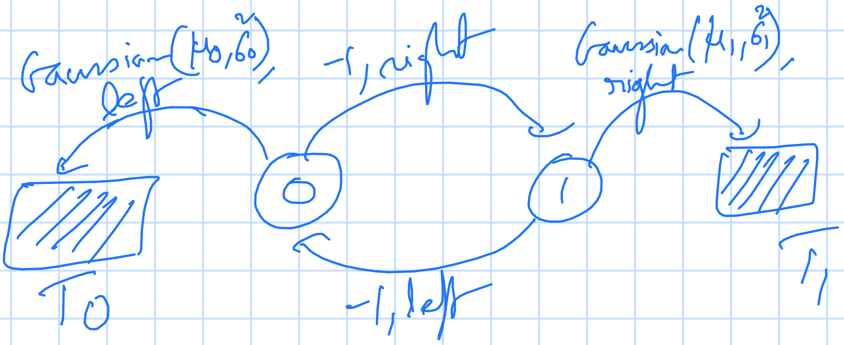
For this, we must weigh the returns by the importance sampling factor. The IS factor has μ_* terms in the numerator & corresponding μ terms in the denominator.

Further observe that for every return, the episode (sub-)sequence to the terminal state has at least one action which is chosen by μ_* with probability 0.

As a result, the estimates $V_*(0)$ & $V_*(1)$ are:

$$V_*(0) = 0.$$
$$V_*(1) = 0.$$

Question 4. 30 **marks** Consider the MDP in Figure 1b. As always, we want the agent to take actions that maximize expected return. Also, we want to structure the rewards such that the actions that maximize the return have the agent end in terminal state T_1, starting from either 0 or 1. Provide conditions on $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$ that will satisfy our requirements. Assume that the discounting factor γ can take any value in $(0, 1]$.



$$V_\pi(0) = \max\{-1 + \gamma V_\pi(1),\ \mu_0\}$$

$$V_\pi(1) = \max\{\mu_1,\ -1 + \gamma V_\pi(0)\}$$

13

We want the agent to take the expected return maximizing actions, such that they result in the agent moving in T_1.

It suffices to have

$$-1 + \gamma V_\pi(1) > \mu_0$$

and

$$\mu_1 > -1 + \gamma V_\pi(0)$$

13

Set $V_\pi(1) = \mu_1$. [This is the max an agent can get when terminating in T_1, starting in state 1]

We desire:

$$-1 + \gamma\mu_1 > \mu_0$$

$$\mu_1 > \left(\frac{\mu_0 + 1}{\gamma}\right)$$