Reinforcement Learning

End Semester Exam

09/12/2024

Sanjit K. Kaul

Instructions: You have two hours to work on the questions. Answers without supporting steps will receive zero credit. Any resources, other than a pen/pencil, are **not** allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 50 **marks** A robot chooses an action at every decision instant $t \in \{0, 1, 2, \ldots\}$. It has been given a policy PMF π_{base} that captures how a human expert makes decisions. The robot also knows the MDP that describes the environment.

Consider the following k-step lookahead, $k \geq 1$, way of choosing an action. At any time t, given that the robot observes state $S_t = s$, for the sake of calculating action A_t, it *looks ahead* k time steps into the future using the MDP. Specifically, it assumes that actions at time $t + k$ and beyond are chosen using π_{base}. Given this, it calculates the best action in state s at time t. A_t is set to the calculated best action and executed by the robot at t.

Consider an MDP with two non-terminal states 0 and 1 and two terminal states -1 and 2. In states 0 and 1, the robot can choose from the set $\{\text{stay}, \text{change}\}$ of actions. Choosing change in 0 transitions the environment to 1 and the agent receives a reward drawn from the Gaussian distribution with mean 2 and variance 100. Choosing change in 1 transitions the environment to 0 and the agent receives a reward drawn from the Gaussian distribution with mean 2 and variance 1. Choosing stay in 0 transitions the environment back to 0 with probability 0.4 and to -1 otherwise. The robot receives a reward drawn from a Gaussian with mean 2 and variance 10 in case the environment transitions to 0. Else, the robot receives a reward of 4. Choosing stay in 1 transitions the environment to 1 with probability 0.4 and to 2 otherwise. The robot receives a reward drawn from a Gaussian with mean 2 and variance 10 in case the environment transitions to 1. Else, the robot receives a reward of 4.

Assume π_{base} chooses actions with equal probability. Let $\gamma = 0.8$. Answer the following questions.

(a) Derive the actions the agent chooses in states 0 and 1 when using 1-step lookahead.

(b) Derive the actions the agent chooses in states 0 and 1 when using 2-step lookahead. [Hint: Think in terms of the Bellman optimality principle.] Does the 2-step lookahead improve upon 1-step lookahead?

(c) Derive the actions the agent chooses in states 0 and 1 when using the optimal policy. How do the lookahead based action selections above compare with the optimal policy?

Question 2. 20 **marks** Consider a MDP with three non-terminal states $0, 1, 2$ and a terminal state -1. In state 0, an agent can either choose to go left or go right. Choosing left has the environment transition to state -1. Choosing right transitions the state to 1. In state 1, an agent can only choose right. This transitions the state to 2. In state 2, the agent can only choose right. This transitions the state to -1. All transitions result in a reward of 1. Consider a policy π_θ, parametrized by the vector θ. Derive the gradient of $v_{\pi_\theta}(0)$ with respect to θ, in terms of the gradient of the policy π_θ and the action-value function q_{π_θ}.

Question 3. 20 **marks** Come up with an approximation architecture for an action-value function approximation, parametrized by the weight vector ω, whose gradient with respect to ω is $(1/\pi(a|s, \theta))\nabla_\theta \pi(a|s, \theta)$, for every state action pair (s, a). Here π is the policy parametrized by θ. For your choice of approximation architecture, derive the ω that results in the best action-value function approximation. Clearly state the dimensions of ω and θ and any other quantities that appear in the process of deriving the best ω.

Question 4. 10 **marks** We are given a Gaussian policy PDF $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/(2\sigma^2)}$, $x \in (-\infty, \infty)$, parametrized by its mean μ and standard deviation σ. Let $\theta = [\mu \ \sigma]^T$. Consider the random variable $A(x) = \sum_{i=1}^{2} a_i \frac{\partial \log p(x)}{\partial \theta_i}$. Derive its inner product with itself (that is, calculate its variance).

Question 1. 50 **marks** A robot chooses an action at every decision instant $t \in \{0, 1, 2, \ldots\}$. It has been given a policy PMF π_{base} that captures how a human expert makes decisions. The robot also knows the MDP that describes the environment.

Consider the following k-step lookahead, $k \geq 1$, way of choosing an action. At any time t, given that the robot observes state $S_t = s$, for the sake of calculating action A_t, it *looks ahead* k time steps into the future using the MDP. Specifically, it assumes that actions at time $t+k$ and beyond are chosen using π_{base}. Given this, it calculates the best action in state s at time t. A_t is set to the calculated best action and executed by the robot at t.

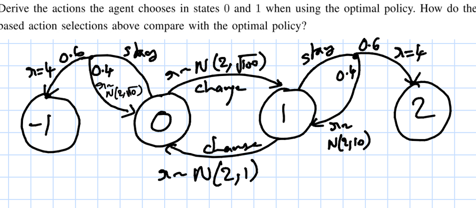
Consider an MDP with two non-terminal states 0 and 1 and two terminal states -1 and 2. In states 0 and 1, the robot can choose from the set $\{\text{stay}, \text{change}\}$ of actions. Choosing change in 0 transitions the environment to 1 and the agent receives a reward drawn from the Gaussian distribution with mean 2 and variance 100. Choosing change in 1 transitions the environment to 0 and the agent receives a reward drawn from the Gaussian distribution with mean 2 and variance 1. Choosing stay in 0 transitions the environment back to 0 with probability 0.4 and to -1 otherwise. The robot receives a reward drawn from a Gaussian with mean 2 and variance 10 in case the environment transitions to 0. Else, the robot receives a reward of 4. Choosing stay in 1 transitions the environment to 1 with probability 0.4 and to 2 otherwise. The robot receives a reward drawn from a Gaussian with mean 2 and variance 10 in case the environment transitions to 1. Else, the robot receives a reward of 4.

Assume π_{base} chooses actions with equal probability. Let $\gamma = 0.8$. Answer the following questions.

(a) Derive the actions the agent chooses in states 0 and 1 when using 1-step lookahead.

(b) Derive the actions the agent chooses in states 0 and 1 when using 2-step lookahead. [Hint: Think in terms of the Bellman optimality principle.] Does the 2-step lookahead improve upon 1-step lookahead?

(c) Derive the actions the agent chooses in states 0 and 1 when using the optimal policy. How do the lookahead based action selections above compare with the optimal policy?



} MDP Up to 5/50.

(a) 1-step lookahead policy π:

$$\pi(0) = \underset{a \in \{\text{stay}, \text{change}\}}{\operatorname{argmax}} E\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+1}) \mid S_t = 0, A_t = a\right]$$

We need to calculate $V_{\pi_{base}}(s)$ for $s = 0, 1$.

$$V_{\pi_{base}}(0) = E_{\pi_{base}}\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+1}) \mid S_t = 0\right]$$

$$= E_{\pi_{base}}\left[R_{t+1} \mid S_t = 0\right] + \gamma E_{\pi_{base}}\left[V_{\pi_{base}}(S_{t+1}) \mid S_t = 0\right]$$

$$= \frac{0.4}{2}(2) + \left(\frac{0.6}{2}\right)4 + \frac{1}{2}(2) + (0.8)\frac{(0.4)}{2}V_{\pi_{base}}(0) + (0.8)\left(\frac{1}{2}\right)V_{\pi_{base}}(1) + \underbrace{(0.8)\frac{(0.6)}{2}(0)}_{\text{corresponding to terminal state } -1.}$$

$$= 2.6 + 0.16\, V_{\pi_{base}}(0) + 0.4\, V_{\pi_{base}}(1).$$

$$\Rightarrow 0.84\, V_{\pi_{base}}(0) = 2.6 + 0.4\, V_{\pi_{base}}(1).$$

Given the symmetry of the MDP and π_{base}, we can write

$$0.84\, V_{\pi_{base}}(1) = 2.6 + 0.4\, V_{\pi_{base}}(0)$$

$$\Rightarrow -0.4\, V_{\pi_{base}}(0) = -0.84\, V_{\pi_{base}}(1) + 2.6$$

$$\Rightarrow \frac{2.6}{0.84} + \frac{0.4}{0.84}V_{\pi_{base}}(1) = \frac{0.84}{0.4}V_{\pi_{base}}(1) - \frac{2.6}{0.4}$$

$$\Rightarrow \frac{2.6}{0.84} + \frac{2.6}{0.4} = V_{\pi_{base}}(1)\left(\frac{0.84}{0.4} - \frac{0.4}{0.84}\right)$$

$$(2.6)(1.24) = V_{\pi_{base}}(1)(0.44)(1.24)$$

$$V_{\pi_{base}}(1) = \frac{2.6}{0.44} = \frac{1.3}{0.22}.$$

Also, $V_{\pi_{base}}(0) = \frac{1.3}{0.22}$. //

} Correct approach & answer (choice of action): Up to 15/50

Correct approach & wrong answer. Up to 12/50

Incorrect approach: Up to 5/50

Now consider $\pi(0)$ & $\pi(1)$, where π is the 1-step lookahead policy. As we said before,

$$\pi(0) = \underset{a \in \{\text{stay}, \text{change}\}}{\operatorname{argmax}} E\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+1}) \mid S_t = 0, A_t = a\right]$$

$$E\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+1}) \mid S_t = 0, A_t = \text{stay}\right]$$

$$= E\left[R_{t+1} \mid S_t = 0, A_t = \text{stay}\right] + \gamma(0.4)V_{\pi_{base}}(0) + \gamma(0.6)V_{\pi_{base}}(-1)$$

$$= (0.4)2 + (0.6)4 + (0.8)(0.4)V_{\pi_{base}}(0)$$

$$= 0.8 + 2.4 + (0.8)(0.4)\frac{(1.3)}{0.22}$$

$$= 3.2 + \frac{0.13}{0.22}(3.2) = 3.2\left(\frac{0.35}{0.22}\right) \quad \text{— (1)}$$

Now consider

$$E\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+1}) \mid S_t = 0, A_t = \text{change}\right]$$

$$= 2 + (0.8)V_{\pi_{base}}(1)$$

$$= 2 + (0.8)\left(\frac{1.3}{0.22}\right). \quad \text{— (2)}$$

Comparing (1) & (2),

$\pi(0) = \text{change}$.

Given the symmetric MDP

$\pi(1) = \text{change}$.

One-step lookahead would choose change in either state.

(b) 2-step lookahead policy μ.

At $t+2$ and beyond we use π_{base}.

Therefore, the value of any state at $t+2$, $E[G_{t+2} \mid S_{t+2} = s] = V_{\pi_{base}}(s)$.

Given this, at any state s that is observed at $t+1$, we must choose the optimal action. That is at $t+1$, we need the value of 1-step lookahead, where we use policy π at $t+1$ & $t+2$ onward we use π_{base}.

Consider state 0:

The value of state 0, time $t+1$ onward is

$$E_{\pi}\left[R_{t+1} + \gamma V_{\pi_{base}}(S_{t+2}) \mid S_{t+1} = 0\right]$$

$$= 2 + (0.8)V_{\pi_{base}}(1)$$

[Because at $t+1$ we use π, which chooses change in 0].

$$= 2 + (0.8)\left(\frac{1.3}{0.22}\right)$$

The corresponding value of state 1 is also $2 + (0.8)\left(\frac{1.3}{0.22}\right)$.

} Correct approach & answer (choice of action) & conclusion Up to 15/50

Correct approach & wrong answer/conclusion Up to 12/50

Incorrect approach: Up to 5/50

[We calculated these values in part (a) & it is okay to just state the values with proper reasoning, without recalculation.]

Given the above values of states 0 & 1 at $t+1$, we must choose the best action at time t.

The 2-step lookahead action selection at time t is

$$\mu(0) = \underset{a \in \{\text{stay}, \text{change}\}}{\operatorname{argmax}} E\left[R_{t+1} + \gamma V_{1\text{-step}}(S_{t+1}) \mid S_t = 0, A_t = a\right]$$

Consider:

$$E\left[R_{t+1} + \gamma V_{1\text{-step}}(S_{t+1}) \mid S_t = 0, A_t = \text{stay}\right]$$

$$= (0.6)4 + (0.4)(2) + (0.8)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right)$$

$$= 3.2 + (0.8)(0.4)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right) \quad \text{— (3)}$$

Further,

$$E\left[R_{t+1} + \gamma V_{1\text{-step}}(S_{t+1}) \mid S_t = 0, A_t = \text{change}\right]$$

$$= 2 + (0.8)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right) \quad \text{— (4)}$$

$$= 2 + (0.8)(0.4)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right) + \underbrace{(0.8)(0.6)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right)}_{> 1.2.}$$

Comparing (3) & (4)

$\mu(0) = \text{change}$

Also, $\mu(1) = \text{change}$

Using 1-step the states have value $2 + (0.8)\left(\frac{1.3}{0.22}\right)$.

Using 2-step the value is $2 + (0.8)\left(2 + (0.8)\left(\frac{1.3}{0.22}\right)\right)$

$$= 2 + (0.8)\left(2 + \frac{1.3}{0.22} - \frac{0.26}{0.22}\right) //$$

which is greater than the 1-step value above.

This is not surprising as 2-step is essentially 2 steps of value iteration, starting with $V_{\pi_{base}}(s)$. Unless just one iteration had resulted in the optimal value function, one would expect 2-step to result in a larger value.

Of course, 2-step chooses the optimal policy at both t and $t+1$, that is for one more step in the future than 1-step. So one would expect 2-step to do better, or at least as well as 1-step.

(c) Optimal policy:

} Same as for (b).

Start with the Bellman optimality equations:

$$V_*(0) = \max_{a \in \{\text{stay}, \text{change}\}} E\left[R_{t+1} + \gamma V_*(S_{t+1}) \mid S_t = 0\right]$$

$$= \max\left\{(0.6)4 + (0.4)2 + (0.8)(0.4)V_*(0),\ 2 + (0.8)V_*(1)\right\}$$

$$= \max\left\{3.2 + (0.8)(0.4)V_*(0),\ 2 + 0.8\, V_*(1)\right\}$$

Nothing distinguishes the states 0 & 1. We would expect them to have the same optimal value. That is $V_*(0) = V_*(1) = V_*$, where V_* is unknown.

We have

$$V_* = \max\left\{3.2 + (0.8)(0.4)V_*,\ 2 + 0.8\, V_*\right\}$$

$$= \max\left\{3.2 + 0.32\, V_*,\ 2 + 0.8\, V_*\right\}$$

V_* is maximized by setting

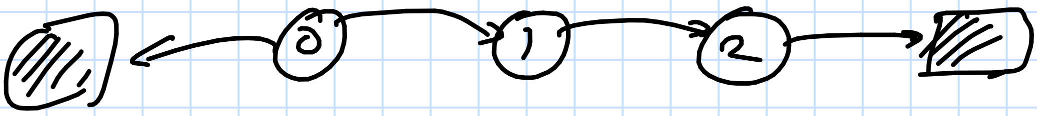
$$V_* = 2 + 0.8\, V_*,$$

which gives $V_* = 10$.

The action change is optimal!

The 2-step value is smaller. /

Question 2. 20 **marks** Consider a MDP with three non-terminal states $0, 1, 2$ and a terminal state -1. In state 0, an agent can either choose to go left or go right. Choosing left has the environment transition to state -1. Choosing right transitions the state to 1. In state 1, an agent can only choose right. This transitions the state to 2. In state 2, the agent can only choose right. This transitions the state to -1. All transitions result in a reward of 1. Consider a policy π_θ, parametrized by the vector θ. Derive the gradient of $v_{\pi_\theta}(0)$ with respect to θ, in terms of the gradient of the policy π_θ and the action-value function q_{π_θ}.



$$v_{\pi_\theta}(0) = \sum_a q_{\pi_\theta}(0,a)\,\pi_\theta(a|0)$$

$$= q_{\pi_\theta}(0,\text{left})\,\pi_\theta(\text{left}|0) + q_{\pi_\theta}(0,\text{right})\,\pi_\theta(\text{right}|0).$$

Correct expansion MDP etc. Up to 5/20.

$q_{\pi_\theta}(0,\text{left})$ is a number (the reward) independent of the policy. Gradient is 0.

Up to 5/20

$\nabla_\theta\, q_{\pi_\theta}(0,\text{right})$

$$= \nabla_\theta\left[r + q_{\pi_\theta}(1,\text{right})\right]$$

$$= \nabla_\theta\left[q_{\pi_\theta}(1,\text{right})\right] = \nabla_\theta\left[r + q_{\pi_\theta}(2,\text{right})\right] = 0.$$

Up to 5/20

Therefore:

$$\nabla_\theta v_{\pi_\theta}(0) = \left[q_{\pi_\theta}(0,\text{left})\,\nabla\pi_\theta(\text{left}|0) + q_{\pi_\theta}(0,\text{right})\,\nabla\pi_\theta(\text{right}|0)\right]$$

Final expression: Up to 5/20

Question 3. 20 marks Come up with an approximation architecture for an action-value function approximation, parametrized by the weight vector ω, whose gradient with respect to ω is $(1/\pi(a|s,\theta))\nabla_\theta \pi(a|s,\theta)$, for every state action pair (s,a). Here π is the policy parametrized by θ. For your choice of approximation architecture, derive the ω that results in the best action-value function approximation. Clearly state the dimensions of ω and θ and any other quantities that appear in the process of deriving the best ω.

A linear approximation architecture that approximates as:

$$Q_\pi(s,a) \approx \frac{1}{\pi(a|s,\theta)} \nabla_\theta \pi(a|s,\theta)^T \omega.$$

Let ω be a $n\times 1$ column vector. The gradient vector must have the same dimensions.

Approx arch. Up to: 5/20

We would like to minimize the mean squared approximation error, average over all (s,a).

The error is:

$$\sum_{s,a} \rho^\pi(s)\,\pi(a|s,\theta)\left(\frac{1}{\pi(a|s,\theta)}\nabla_\theta \pi(a|s,\theta)^T \omega - Q_\pi(s,a)\right)^2.$$

We require the gradient of the error w.r.t ω to be 0 at the ω that minimizes the error.

Identifying & stating the error we want to minimize. Up to: 5/20

The gradient

$$\sum_{s,a} \rho^\pi(s)\,\pi(a|s,\theta)\left(\frac{1}{\pi(a|s,\theta)}\nabla_\theta \pi(a|s,\theta)^T \omega - Q_\pi(s,a)\right)\frac{1}{\pi(a|s,\theta)}\nabla_\theta \pi(a|s,\theta) = 0$$

$$\Rightarrow \sum_{s,a} \rho^\pi(s)\left(\underbrace{\nabla_\theta \pi(a|s,\theta)^T \omega}_{\text{scalar}}\ \underbrace{\nabla_\theta \pi(a|s,\theta)}_{\text{vector}} - Q_\pi(s,a)\nabla_\theta \pi(a|s,\theta)\right) = 0$$

$$\sum_{s,a} \rho^\pi(s)\,\nabla_\theta \pi(a|s,\theta)\,\nabla_\theta \pi(a|s,\theta)^T \omega = \sum_{s,a} \rho^\pi(s)\,Q_\pi(s,a)\,\nabla_\theta \pi(a|s,\theta)$$

The optimal ω is

$$\omega^* = \left(\sum_{s,a} \rho^\pi(s)\,\nabla_\theta \pi(a|s,\theta)\,\nabla_\theta \pi(a|s,\theta)^T\right)^{-1} \sum_{s,a} \rho^\pi(s)\,Q_\pi(s,a)\,\nabla_\theta \pi(a|s,\theta)$$

Setting the derivative 0 & the answer. Up to: 10/20

Question 4. 10 **marks** We are given a Gaussian policy PDF $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/(2\sigma^2)}$, $x \in (-\infty, \infty)$, parametrized by its mean μ and standard deviation σ. Let $\theta = [\mu \ \sigma]^T$. Consider the random variable $A(x) = \sum_{i=1}^{2} a_i \frac{\partial \log p(x)}{\partial \theta_i}$. Derive its inner product with itself (that is, calculate its variance).

$$\mathrm{Var}[A(x)] = \mathrm{Corr}[A(x), A(x)] - \left(E[A(x)]\right)^2$$

$$E[A(x)]$$
$$= E\left[\sum_{i=1}^{2} a_i \frac{\partial}{\partial \theta_i} \log p(x)\right]$$
$$= \sum_{i=1}^{2} a_i \, E\left[\frac{\partial}{\partial \theta_i} \log p(x)\right].$$

$$E\left[\frac{\partial}{\partial \theta_i} \log p(x)\right]$$
$$= \int \left(\frac{\partial}{\partial \theta_i} \log p(x)\right) p(x)\, dx$$
$$= \int \frac{1}{p(x)} \frac{\partial p(x)}{\partial \theta_i} p(x)\, dx$$
$$= \int \frac{\partial p(x)}{\partial \theta_i} dx = \frac{\partial}{\partial \theta_i} \int p(x)\, dx = 0.$$

Up to 5/10

$$\therefore \mathrm{Var}[A(x)] = \mathrm{Corr}[A(x), A(x)]$$
$$= E[A(x) A(x)]$$
$$= E\left[\left(\sum_{i=1}^{2} a_i \frac{\partial}{\partial \theta_i} \log p(x)\right)\left(\sum_{j=1}^{2} a_j \frac{\partial}{\partial \theta_j} \log p(x)\right)\right]$$
$$= E\left[\sum_{i=1}^{2}\sum_{j=1}^{2} a_i a_j \frac{\partial}{\partial \theta_i} \log p(x) \frac{\partial}{\partial \theta_j} \log p(x)\right]$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} a_i a_j \, E\left[\frac{\partial}{\partial \theta_i} \log p(x) \frac{\partial}{\partial \theta_j} \log p(x)\right]$$

$$E\left[\frac{\partial}{\partial \theta_i} \log p(x) \frac{\partial}{\partial \theta_j} \log p(x)\right]$$
$$= E\left[\frac{1}{p(x)} \frac{\partial p(x)}{\partial \theta_i} \frac{1}{p(x)} \frac{\partial p(x)}{\partial \theta_j}\right]$$
$$= \int_{-\infty}^{\infty} \frac{1}{p(x)} \frac{1}{p(x)} \frac{\partial p(x)}{\partial \theta_i} \frac{\partial p(x)}{\partial \theta_j} p(x)\, dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{p(x)} \frac{\partial p(x)}{\partial \theta_i} \frac{\partial p(x)}{\partial \theta_j} dx.$$

Up to 8/10

Note $\theta_1 = \mu$, $\theta_2 = \sigma$.

$$\frac{\partial}{\partial \mu}\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \left(\frac{+2(x-\mu)}{2\sigma^2}\right)$$
$$= \frac{(x-\mu)}{\sigma^2\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Up to 9/10

$$\frac{\partial}{\partial \sigma}\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\right)$$
$$= \frac{-1}{\sigma^2\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} + \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-(x-\mu)^2/(2\sigma^2)}\right)\left(\frac{(x-\mu)^2}{\sigma^3}\right)$$

Up to 10/10.

Differences that result from not correcting the typo in the definition of $p(x)$ in the question should be ignored.