Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE675A)
End Semester, Date: 1st May, 2023

Timing: 5:30 to 8:30 PM | 2022-23 Sem 2 | Max mark: 35 + 5 (Bonus)

I Bandit Algorithms (7 Marks)

1. Answer the following TRUE/FALSE questions related to the *successive elimination* bandit algorithm. No justification is required. [**3 marks**]

   (a) The arm with the smallest LCB gets eliminated in that round. False

   (b) The arm with the smallest UCB gets eliminated in that round. False

   (c) At least one arm gets eliminated in each round. False

   (d) At most one arm gets eliminated in each round. False

   (e) Consider two arms a_1 and a_2 such that $UCB(a_1) < UCB(a_2)$. If a_2 gets eliminated, it implies that a_1 also gets eliminated in that round. True

   (f) Consider two arms a_1 and a_2 such that $LCB(a_1) < LCB(a_2)$. If a_2 gets eliminated, it implies that a_1 also gets eliminated in that round. False

2. If we know that the true means of all the arms are very close to each other, which regret bound will give a better upper bound? Instance-independent bound or Instance-dependent bound? Explain. [**2 marks**]

   Instance dependent bound $O(\frac{klogT}{\Delta_{min}})$
   Instance independent bound
   O($\sqrt{kTlogT}$)
   If true means are all very close, Δ_{min} will be small and blow up the bound and give a bad bound in comparison to the instance independent bound.

   Hence Instance independent bound would be better.

3. Consider solving a 2-arm bandit problem with Thompson sampling. Let D_a and D_b be the posterior distributions corresponding to the two arms a and b at some time t. If $E[D_a] > E[D_b]$, then which of the following statements is correct? Justify your answer. [**2 marks**]

   (a) Arm a will be played next

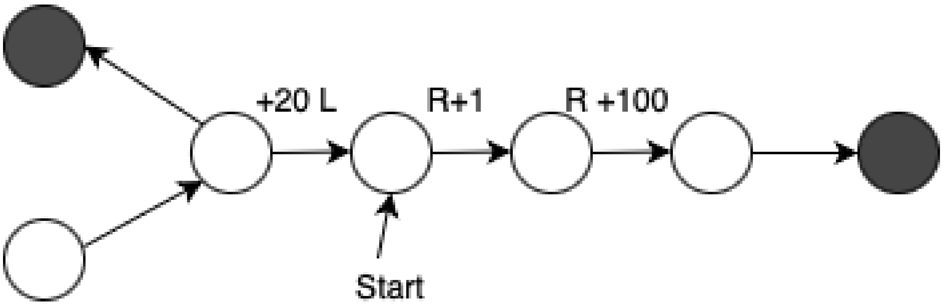
(b) Arm b will be played next

(c) Either of them could be played next

(c) Either
In Thomspon sampling we sample from D$_a, D_b$;$\theta_a \sim D_a, \theta_b \sim D_b$
If $\theta_a > \theta_b$ then we play a.
Even if $E[D_a] > E[D_b]$
There is non-zero probability that $\theta_a < \theta_b$ [vice versa]

II MDP Basics and Dynamic Programming (9 Marks)

4. (*True* or *False*) If the only difference between two MDPs is the value of the discount factor then they must have the same optimal policy. Justify your answer. **[2 marks]** False



Consider the MDP
If γ is high then right would be optimal.
If γ is low then left would be optimal.

5. (*True* or *False*) Policies found by value iteration are superior to policies found by policy iteration. No justification is required. **[1 marks]** False

6. We know that the value iteration update is given by

$$v_{k+1}(s) = \max_a R_s^a + \gamma \sum_{s'} v_k(s') P_{ss'}^a.$$

If we want to design value iteration on action-values $Q(s,a)$ instead of state values $v(s)$, what is the corresponding value iteration update equation for $Q_{k+1}(s,a)$ in terms of $Q_k(s,a)$? **[2 marks]**

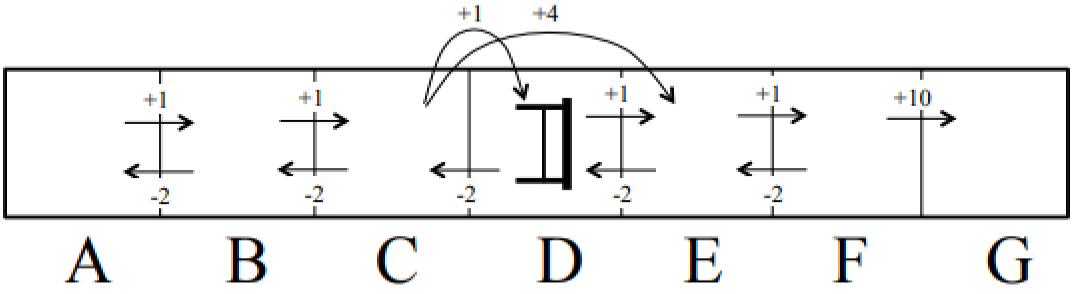
Bellman's optimality equation for $q_k(s,a)$

$$q_k(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} q_k(s',a')$$

Thus

$$q_{k+1}(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} q_k(s',a')$$

7. Consider an MDP modeling a hurdle race track, shown below. There is a single hurdle in square D. The terminal state is G. The agent can run *left* or *right*. If the agent is in square C, it cannot run *right*. Instead, it can *jump*, which either results in a fall to the hurdle square D or a successful hurdle jump to square E both of which are equally likely. Rewards are shown in the figure below. Assume a discount of $\gamma = 1$.



Actions:

- *right:* Deterministically move to the right.
- *left:* Deterministically move to the left.
- *jump:* Stochastically jump to the right. This action is available for square C only.

(a) For the policy π of always moving forward (i.e., using actions *right* or *jump*, whichever is applicable), compute $V^\pi(C)$. **[1 marks]**
$V_\pi(F) = 10$
$V_\pi(E) = 1 + V_\pi(F) = 11$
$V_\pi(D) = 1 + V_\pi(E) = 12$
$V_\pi(C) = R_s^\pi + \gamma \sum_{s'} P_{ss'}^\pi v_\pi(s')$=$4+1\frac{}{2}+\frac{1}{2}$ $\times 11$ + $\frac{1}{2}$ $\times 12$ = $\frac{28}{2}$= 14

(b) Perform two iterations of value iteration and write the obtained estimate for the value function $V_2(s)$. Assume that the initial values $V_0(s) = 0, \forall s$. [**2 marks**]

$$v_{k+1}(s) = \max_a R_s^a + \gamma \sum_{s'} v_k(s') P_{ss'}^a.$$

$$\begin{aligned} V_1(F) &= 10 & V_2(F) &= 10 \\ V_1(E) &= 1 & V_2(E) &= 1 + 10 = 11 \\ V_1(D) &= 1 & V_2(D) &= 1 + 1 = 2 \\ V_1(C) &= 2.5 & V_2(C) &= 2.5 + \frac{1}{2} + \frac{1}{2} = 3.5 \\ V_1(B) &= 1 & V_2(B) &= 1 + 2.5 = 3.5 \\ V_1(A) &= 1 & V_2(A) &= 1 + 1 = 2 \end{aligned}$$

(c) Is that estimate $V_2(s)$ the optimal value function? [**1 marks**]
No it is not.
Since $V_2(c) < V_\pi(c)$ (obtained in part (a)), it is clearly not the optimal value function.

III Model-free methods and Function approximation (12 Marks)

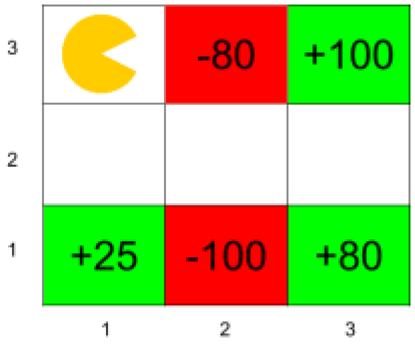
8. (*True* or *False*) When using features to represent the Q-function it is guaranteed that this feature-based Q-learning finds the same Q-function, Q^*, as would be found when using a tabular representation for the Q-function. Justify. [**1 marks**]
False
When we start using features to represent Q-functions, we could get the effect of partially observed MDPs. Multiple (state-action) pairs may have the same representation. Hence it is only an approximation and is not necessary that it would converge to the same optimal Q-function as tabular methods.

9. Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1 - p$. Let the reward be +1 on all transitions, and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps, with a return of 10 . What are

the first-visit and every-visit estimators of the value of the nonterminal state? **[2 marks]**

MC - First visit
return is +10
MC - Every visit
$\frac{10+9+8+...+1}{10} = \frac{10\times 11}{2} \times \frac{1}{10} = 5.5$

10. Consider the grid-world given below and Pacman who is trying to learn the optimal policy. If an action results in landing into one of the shaded states the corresponding reward is awarded during that transition. All shaded states are terminal states, i.e., the MDP terminates once arrived in a shaded state. The other states have the North, East, South, West actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grad). Assume the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state (1,3), i.e., the top left corner.



(a) What is the value of the optimal value function V^* at the following states: **[2 marks]**

$$V^*(3,2) = \dots \quad V^*(2,2) = \dots \quad V^*(1,3) = \dots$$

(b) The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r).

Episode 1	Episode 2	Episode 3
$(1,3), \mathrm{S}, (1,2), 0$	$(1,3), \mathrm{S}, (1,2), 0$	$(1,3), \mathrm{S}, (1,2), 0$
$(1,2), \mathrm{E}, (2,2), 0$	$(1,2), \mathrm{E}, (2,2), 0$	$(1,2), \mathrm{E}, (2,2), 0$
$(2,2), \mathrm{S}, (2,1), -100$	$(2,2), \mathrm{E}, (3,2), 0$	$(2,2), \mathrm{E}, (3,2), 0$
	$(3,2), \mathrm{N}, (3,3), +100$	$(3,2), \mathrm{S}, (3,1), +80$

Using Q-Learning updates, what are the following Q-values after the above three episodes. Assume Q-values are initialized to zeros. **[2 marks]**

$$Q((3,2),\mathrm{N}) = \dots \quad Q((1,2),\mathrm{S}) = \dots \quad Q((2,2),\mathrm{E}) = \dots$$

(c) Consider a feature-based representation of the Q-value function:

$$Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$f_1(s)$: The x coordinate of the state $\quad f_2(s)$: The y coordinate of the state

$$f_3(N) = 1, f_3(S) = 2, f_3(E) = 3, f_3(W) = 4$$

1. Given that all w_i are initially 0 , what are their values after the first episode? **[2 marks]**
2. Assume the weight vector w is equal to $(1,1,1)$. What is the action prescribed by the Q-function in state $(2,2)$ **[1 marks]**?

(a) $V^*(3,2) = +100$, $V^*(2,2) = +50$, $V^*(1,2) = +25$, $V^*(1,3) = +12.5$

(b) Q((1,3),S) $= 0 + \frac{1}{2} \times [0.5 \times 0 - 0] = 0$
Q((1,2),E) $= 0$
Q((2,2),S) $= 0 + \frac{1}{2}[-100 + 0 - 0] = -50$
Q((1,3),S) $= 0$
Q((1,3),E) $= 0 + \frac{1}{2}[0 + \frac{1}{2}[0] - 0] = 0$
Q((2,2),E) $= 0 + \frac{1}{2}[0 + \frac{1}{2}[0] - 0] = 0$
Q((3,2),N) $= 0 + \frac{1}{2}[100 + \frac{1}{2}[0] - 0] = +50$
Q((1,3),S) $= 0 + \frac{1}{2}[0 + \frac{1}{2}[0] - 0] = 0$
Q((1,2),E) $= 0 + \frac{1}{2}[0 + \frac{1}{2}[0] - 0] = 0$
Q((2,2),E) $= 0 + \frac{1}{2}[0 + \frac{1}{2}[50] - 0] = 0 + 12.5 = 12.5$
Q((3,2),S) $= 0 + \frac{1}{2}[80 + [0] - 0] = 40$
Q((3,2),N) $= +50$
Q((1,2),S) $= 0$
Q((2,2),E) $= 12.5$

c) We need $|Q_f(s,a;w) - Q(s,a)|^2$ to be minimized
1) $w_{t+1} = w_t + \alpha[q^*(s_t, A_t) - \hat{q}(s_t, A_t; w_t)]\nabla\hat{q}(s_t, A_t; w_t)$
$w_1 = w_0 + \frac{1}{2}[0 + 0 - 0] \times [...]^T = w_0$
$w_2 = w_1 + \frac{1}{2}[0 + 0 - 0] \times [...]^T = w_1$
$w_3 = w_2 + \frac{1}{2}[-100 + 0 - 0] \times [2,2,2]^T = [-100, -100, -100]^T$

2) Q((2,2),N) $= 5$

Q((2,2),S) = 6
Q((2,2),E) = 7
Q((2,2),W) = 8
Hence w is the optimal action at (2,2) prescribed by the Q function.

a) Yes, it will converge to the optimal policy. Under appropriate conditions, if all states and actions are visited often, then even though behavior is different from target policy, it will still converge.

b) No, it will not converge to the optimal Q value function. Instead it will converge to the Q value function of a policy that is ϵ-close to $\pi[\epsilon = 0.5]$

11. Recall that reinforcement learning agents gather tuples of the form $< s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1} >$ to update the value or Q-value function. In both of the following cases, the agent acts at each step as follows: with probability 0.5 it follows a fixed (not necessarily optimal) policy π and otherwise it chooses an action uniformly at random. Assume that in both cases updates are applied infinitely often, state-action pairs are all visited infinitely often, the discount factor satisfies $0 < \gamma < 1$, and learning rates α are all decreased at an appropriate pace.

    (a) Agent-1 performs the following update:

    $$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\left[r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right]$$

    Will this process converge to the optimal Q-value function? If yes, write "Yes." If not, give an interpretation (in terms of kind of value, optimality, etc.) of what it will converge to, or write that it will not converge. **[1 marks]**

    Yes, Q-learning just needs infinite exploration in behavioural policy.

    (b) Agent-2 performs the update

    $$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\left[r_{t+1} + \gamma Q\left(s_{t+1}, a_{t+1}\right) - Q\left(s_t, a_t\right)\right]$$

    Will this process converge to the optimal Q-value function? If yes, write "Yes." If not, give an interpretation (in terms of kind of value, optimality, etc.) of what it will converge to, or write that it will not converge. **[1 marks]**

    Converges to value function of Behavioural policy being followed.

IV Policy Gradient Methods (7 marks)

12. Consider a parameterized representation of a policy π given by

$$\pi\left(a|s;\theta\right) = \frac{\exp(\theta^T x(s,a))}{\sum\limits_{b\in\mathcal{A}} \exp(\theta^T x(s,b))}, \quad \text{for } a \in \mathcal{A}.$$

$\ln \pi(a|s;\theta) = \theta^T.x(s,a) - \ln(\sum_{b\in A} e^{\theta^T x(s,b)})$

$$\begin{aligned}\nabla_\theta \ln \pi(a|s;\theta) &= x(s,a) - \frac{1}{\sum_{b\in A} e^{\theta^T x(s,b)}}.\sum_{b\in A} e^{\theta^T x(s,b)}.x(s,b)\\ &= x(s,a) - \sum_{b\in A} \frac{e^{\theta^T x(s,b)}}{\sum_{a'\in A} e^{\theta^T x(s,a')}}.x(s,b)\\ &= x(s,a) - \sum_{b\in A} \pi(b|s;\theta).x(s,b)\\ &= x(s,a) - \sum_{a'\in A} \pi(a'|s;\theta).x(s,a')\end{aligned}$$

The gradient is approximated

$$E_\pi[R_t \nabla_\theta \log \pi(A_t;\theta)] \approx R_t.\nabla_\theta \log \pi(A_t;\theta)$$

Thus, the update equation

$$\begin{aligned}\theta_{new} &= \theta_{old} + \alpha R_{t+1} \nabla \log \pi(A_t;\theta)\\ &= \theta_{old} + \alpha R_{t+1}.[x(s_t,A_t) - \sum_{a'\in A} \pi(a'|s_t;\theta_{old}).x(s_t,a')]\end{aligned}$$

Here $\mathcal{A}$ is the set of actions, and θ is the parameter vector that characterizes the policy π, and $x(s,a)$ is the feature vector the pair (s,a). Derive the vanilla policy gradient update for this case. **[3 marks]**

13. Write the Actor-critic algorithm (without baseline) in the full RL setting assuming a policy parameterization as given in Question 12 and a linear function approximation for Q-function as given in Question 10(c). **[4 marks]**

Algorithm parameters: policy -parameter step size α, value function parameter step size β. Initialize policy parameter $\theta^{(0)}$ and value function approximator $w^{(0)}$

```
Loop for each episode:

Initialize S
Loop for each step of episode
Sample $A_t$ from $\pi(a|s_t;\theta^t)$
Observe $R_{t+1}, S_{t+1}$
    $\theta^{t+1} \leftarrow \theta^t + \alpha.(w^t)^T.\hat{x}(s_t, A_t).\nabla_\theta \log(\pi(A_t|s_t)|_{\theta^{(t)}})$
    $w^{t+1} \leftarrow w^t + \beta.[R_{t+1} + \gamma \max_a (w^{(t)})^T \hat{x}(s_{t+1}, a) - (w^{(t)})^T \hat{x}(s_t, A_t)].\hat{x}(s_t, A_t)$
Until terminal state
```

V Bonus (5 Marks)

14. Create one interesting question which can be used in the next year's final exam for this course. Give a detailed solution as well.

    Note: The bonus marks will be used only to award A* grades.