INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
End-Autumn Semester 2024-25

Date of Examination: **26/11/24** Session(FN/AN) **AN** **Duration 3 hrs,** **Marks = 80**

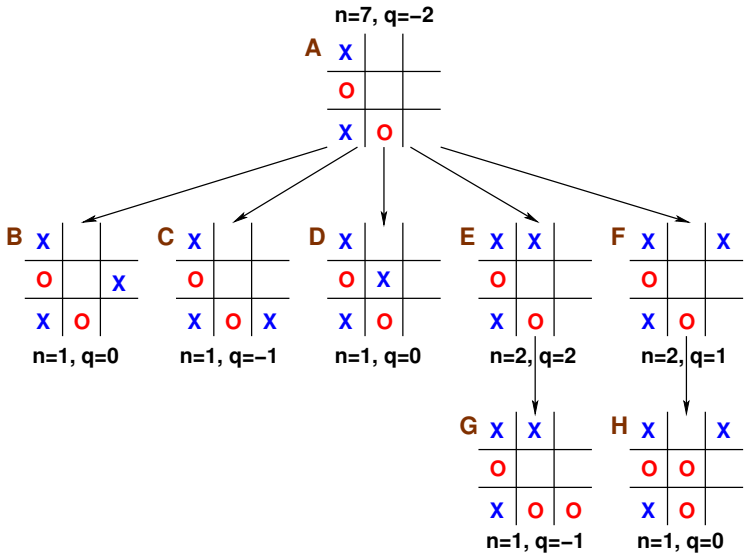
Sub No: **CS60077** Sub Name: **Reinforcement Learning**

Department/Centre/School : **Computer Science and Engineering**

Specific charts, graph paper,log book etc. required **NO**

Special Instructions (if any) In case of reasonable doubt, make practical assumptions and state them upfront. Marks will be deducted for sketchy proofs and claims without proper reasoning. All parts of a single question MUST be done in the same place.

Q1. You are using Monte Carlo Tree Search (MCTS) to decide on the next action for a two-player (Player-'X' and Player-'O') competitive game, called *Tic-Tac-Toe*. So far, the node expansion of the tree from an arbitrary root node/configuration, where Player-'O' makes a move, is shown as follows (each node consists of node identifier/name **A**-**H**, n-value, and q-value).



Recall that, the formula for the UCT value for a node v is:

$$UCT(v) = \frac{q(v)}{n(v)} + c\sqrt{\frac{\ln(\pi(v))}{n(v)}}$$

where, $n(v)$ and $q(v)$ denote the n-value and q-value for node v, and $\pi(v)$ indicates the n-value of the parent node of v. (Assume $c = \frac{1}{2}$ in the UCT formula)

Answer the following questions based on next/one iteration of the MCTS algorithm (Selection/Expansion/Simulation/Backup).

(a) [Selection] What is the node that is next selected? Show your calculations in details. **(5)**

(b) [Expansion] What are the possible configurations that the child node can be expanded into? Show all node configurations as snapshots of the Tic-Tac-Toe board configurations. What will be the initial n-values and q-values for these newly expanded nodes? **(3)**

(c) [Simulation] Assume that, from the possible configurations that the child node can be expanded into (as you discovered in Part (b)), what according to you (apply your human instinct) is the *best move* for Player-'O'? **(1)**

(d) [Backup] Assuming that the simulation (rollout) from the *best choice* (that you made humanly in Part (c)) of expanded node gives finally a value of +1 (that is, Player-'X' wins even your best effort there) as stated above. Show how the backup for that value is calculated to all of the affected nodes. **(3)**

(e) Assume that after this final rollout, we have run out of time to run the MCTS simulation and must now choose an action for Player-'X' from node A. Which move will be chosen by Player 'X' and why? **(2)**

Q2. A robot in a grid world follows a fixed policy. The robot has the option to travel to any neighboring grid but not diagonally, i.e., cannot go to (2, 2) from (1, 1), or cannot go to (2, 1) from (1, 2), or vice-versa. Use Value Iteration to evaluate a policy for a 2x2 grid world, where all transitions yield a reward of -1 except reaching the goal at (2,2) with reward 0.

(a) Write with explanation the Bellman Expectation Equations and Optimality Equations for both value functions $V(s)$ and $Q(s, a)$. **(4)**

(b) Initialize $V(s) = 0$ for all four states $(1, 1), (1, 2), (2, 1), (2, 2)$ and perform one iteration of Value Iteration, updating the value function for each state. Assume $\gamma = 1$. **(4)**

(c) Describe in detail the differences between Policy Evaluation, Policy Iteration, and Value Iteration algorithms with respect to algorithm and goal. Use equations wherever necessary. **(6)**

(d) Explain with suitable motivation the difference in the update rules and neural learning architectures of DQN and Double DQNs. **(6)**

Q3. **(a)** For the Duelling DQN method, answer the following questions.

(i) Explain the motivation behind separating value and advantage functions. **(2)**

(ii) Explain how two streams (value & advantage) are combined to estimate Q-values. **(2)**

(iii) Explain how this reduces noise in action value estimates. **(2)**

(b) Consider a 2x2 grid world with the following states:

- The set of states $\mathcal{S} = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$ with the start state as $(1, 1)$,
- The goal state is $(2, 2)$ with a reward of $+1$,
- All other transitions yield a reward of 0.

The actions available are `Up`, `Down`, `Left`, and `Right`. If an action moves the agent out of the grid boundaries, the agent stays in the same position. Assume a discount factor $\gamma = 0.9$. Assume a policy gradient approach is being used.

(i) Write the mathematical expression of *softmax* based stochastic policy for choosing the actions. **(2)**

(ii) Recall that the gradient of the objective $J(\theta)$ with respect to θ is given by:

$$\nabla_\theta J(\theta) = \sum_t G_t \nabla_\theta \log \pi(a_t|s_t; \theta)$$

Let θ be the parameter vector initialized as follows.

$$\theta = \begin{bmatrix} \theta_{(1,1),right} & \theta_{(1,1),down} & \theta_{(1,1),left} & \theta_{(1,1),up} \\ \theta_{(1,2),right} & \theta_{(1,2),down} & \theta_{(1,2),left} & \theta_{(1,2),up} \\ \theta_{(2,1),right} & \theta_{(2,1),down} & \theta_{(2,1),left} & \theta_{(2,1),up} \\ \theta_{(2,2),right} & \theta_{(2,2),down} & \theta_{(2,2),left} & \theta_{(2,2),up} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Consider that the policy gradient method REINFORCE is used for a single episode from (1, 1) to (2,2), with the action sequence $\{right,\ down\}$ that reaches the goal. Compute and report the updated policy parameters $\theta_{(1,1),right}$ and $\theta_{(1,2),down}$ based on this episode as defined. Consider $\alpha = 0.1$. You must explain all steps and use suitable expressions with standard notations. **(7)**

(c) Clearly state and then prove the "Compatible Function Approximation Theorem". **(5)**

Q4. In Trust Region Policy Optimization (TRPO), the goal is to find a policy π_θ, that maximizes the expected return while ensuring that the updated policy does not deviate too far from current policy.

(a) Considering π_θ and $\pi_{\theta'}$ as two *nearby* policies, prove that

$$J(\theta) - J(\theta') = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_{\theta'}} \\ a\sim\pi_{\theta'}}} [A^{\pi_\theta}(s,a)]$$

where all the symbols have their usual meanings. You can assume the usual definitions of J and A as starting points of the derivation. **(6)**

(b) Starting from the Taylor series expansion, show that the KL divergence constraint can be approximated as:

$$D_{\mathrm{KL}}(\pi_{\theta_{\mathrm{old}}} || \pi_\theta) \approx \frac{1}{2} \Delta\theta^\top \mathbf{F} \Delta\theta$$

where $\mathbf{F}$ is the Fisher information matrix and $\Delta\theta = \theta - \theta_{\mathrm{old}}$. Provide all the necessary reasoning, do not jump steps. **(4)**

(c) Show that the ratio $(\pi_\theta(a|s))/(\pi_{\theta_{\mathrm{old}}}(a|s))$ can be expressed in terms of θ and θ_{old} using the exponential of the policy's log probabilities as given by,

$$\frac{\pi_\theta(a|s)}{\pi_{\theta_{\mathrm{old}}}(a|s)} = \exp\left((\theta - \theta_{\mathrm{old}})^\top \nabla_\theta \log \pi_{\theta_{\mathrm{old}}}(a|s)\right)$$

Provide all the necessary reasoning, do not jump steps. [Hint: $x = \exp(\log x)$ and finally use Taylor series expansion] **(4)**

Q5. For the following questions, multiple answers may be correct. Marks can be obtained only when all correct answers are reported. For each question, mention the correct answer choice(s) with brief justifications/derivations/calculations wherever applicable. **(1×12)**

(a) After 20 iterations of the UCB1 algorithm applied on a 4-arm bandit problem, we have $n_1 = 5,\ n_2 = 4,\ n_3 = 6,\ n_4 = 5$ and $q_1 = 0.58,\ q_2 = 0.61,\ q_3 = 0.65,\ q_4 = 0.55$. Which arm should be played next?

i. `ARM-1`
ii. `ARM-2`
iii. `ARM-3`
iv. `ARM-4`

(b) Using MAXQ approach for hierarchical RL leads to solutions which are:

i. hierarchically optimal
ii. recursively optimal
iii. flat optimal
iv. non-optimal

(c) What is the primary purpose of experience replay in DQN?

i. To improve the efficiency of training
ii. To stabilize the training process.
iii. To reduce the variance of the gradient estimates.
iv. All of the above.

(d) Which of the following expressions correctly represents the policy gradient estimator using the likelihood ratio trick?

i. $\nabla_\theta J(\theta) = \mathbb{E}_\pi[\nabla_\theta \log \pi_\theta(s,a) Q_\pi(s,a)]$
ii. $\nabla_\theta J(\theta) = \mathbb{E}_\pi[\nabla_\theta \pi_\theta(s,a) Q_\pi(s,a)]$
iii. $\nabla_\theta J(\theta) = \mathbb{E}_\pi[\log \pi_\theta(s,a) \nabla_\theta Q_\pi(s,a)]$
iv. $\nabla_\theta J(\theta) = \mathbb{E}_\pi[\pi_\theta(s,a) \nabla_\theta Q_\pi(s,a)]$

(e) In Prioritized Experience Replay, what is the purpose of importance sampling weights?
i. To increase the probability of sampling rare events
ii. To normalize the replay buffer size
iii. To correct the bias introduced by non-uniform sampling
iv. To adjust the learning rate based on priority

(f) What is the difference between Dueling DQN and regular/standard DQN?
i. Dueling DQN uses a separate stream to estimate the state value.
ii. Dueling DQN uses a different loss function.
iii. Dueling DQN uses a different exploration strategy.
iv. Dueling DQN clips Q-values.

(g) The Noisy DQN introduces exploration through:
i. Adding Gaussian noise to the state observations
ii. Random action selection with decreasing probability
iii. Adding parametric noise to the network weights
iv. Randomly dropping network connections

(h) What is the purpose of the advantage function in Actor-Critic methods?
i. To improve the convergence speed of the algorithm.
ii. To reduce the variance of the policy gradient estimate.
iii. To stabilize the training process.
iv. All of the above.

(i) Consider the use of softmax based policy in the policy gradient paradigm of RL algorithms where actions are weighted using a linear combination of features $\phi(s,a)^T\theta$. In that case, the score function will be,
i. $\phi(s,a) - \mathbb{E}_{\pi_\theta}[\phi(s,\cdot)]$
ii. $\phi(s,a)^T\theta$
iii. $\phi(s,a)\theta^T$
iv. $\phi(s,a) + \mathbb{E}_{\pi_\theta}[\phi(s,\cdot)]$

(j) Which of the following techniques can be used to improve the efficiency of policy gradient methods?
i. Natural Policy Gradient
ii. Generalized Advantage Estimation (GAE)
iii. Experience Replay
iv. Importance Sampling

(k) The Distributional DQN differs from regular/standard DQN by:
i. Using a different distribution for exploration
ii. Learning the full distribution of returns instead of just the expected value
iii. Sampling actions from a learned distribution
iv. Applying dropout to create a distribution of Q-values

(l) For Natural Policy Gradient methods, which are true?
i. They are invariant to parameterization of the policy
ii. They always converge faster than vanilla policy gradient
iii. They use the Fisher Information Matrix for step direction
iv. They can only be used with Gaussian policies