Indian Institute of Technology Patna
End-Semester Examination – November 2025
Course: Reinforcement Learning (CS6109/CS603)
Date: 20 November 2025 Time: 1 Hour
Maximum Marks: 50

Name: _____ **Roll No.:** _____

Q. No.	Opt.	Q. No.	Opt.	Q. No.	Opt.	Q. No.	Opt.
1	_____	11	_____	21	_____	31	_____
2	_____	12	_____	22	_____	32	_____
3	_____	13	_____	23	_____	33	_____
4	_____	14	_____	24	_____	34	_____
5	_____	15	_____	25	_____	35	_____
6	_____	16	_____	26	_____	36	_____
7	_____	17	_____	27	_____	37	_____
8	_____	18	_____	28	_____	38	_____
9	_____	19	_____	29	_____	39	_____
10	_____	20	_____	30	_____	40	_____

Multiple Choice Questions

1. Which of the following statements is *correct* ?
   (a) VPG and DQN are both off-policy methods and can only be used in discrete action spaces.
   (b) VPG is on-policy and restricted to discrete action spaces, while DQN is off-policy and can handle both discrete and continuous action spaces.
   (c) PPO is off-policy and restricted to continuous action spaces, while DQN is on-policy and restricted to discrete action spaces.
   (d) VPG and PPO are on-policy methods that can be used with both discrete and continuous action spaces, whereas DQN is an off-policy method restricted to discrete action spaces.

   Correct option: (d)
2. Consider the *exploration* behaviour of Vanilla Policy Gradient (VPG). Which option best captures how exploration typically evolves during training?
   (a) VPG maintains a fixed level of randomness in its policy throughout training to avoid overfitting.
   (b) VPG starts with a stochastic policy, but the policy typically becomes progressively less random as training proceeds.
   (c) VPG initially uses a deterministic policy and gradually adds more stochasticity as training proceeds.
   (d) VPG never changes the degree of randomness; only the value function is updated.

   Correct option: (b)
3. The issue of "overstepping" in Vanilla Policy Gradient (VPG). Which of the following is the most precise reason this overstepping occurs?
   (a) The policy gradient gives both the optimal direction and the optimal step size, but numerical precision errors corrupt the step size.
   (b) The first-order gradient provides no useful information about the policy update direction, only about the step size.

(c) The first-order gradient specifies a direction in which to update the policy but does not specify how far to move, so a fixed learning rate can cause both overshooting and undershooting.

(d) Overstepping in VPG arises because the algorithm is off-policy and therefore uses data from an outdated behaviour policy.

Correct option: (c)

4. According to the description of Proximal Policy Optimization (PPO), what central design question motivates PPO?

(a) How to compute exact second-order derivatives to obtain the Newton step for the policy.

(b) How to choose the best entropy bonus to ensure persistent exploration in a deterministic policy.

(c) How to take the largest possible improvement step on the policy with the available data, without stepping so far that policy performance collapses.

(d) How to convert any off-policy algorithm into an on-policy algorithm by modifying the replay buffer.

Correct option: (c)

5. In the context of PPO , why can clipping be interpreted as a form of regularization?

(a) Because clipping forces the policy parameters to remain within a fixed numerical range, independent of the data.

(b) Because clipping rescales rewards to a bounded interval, stabilizing the learning signal for the critic.

(c) Because clipping limits the change in the policy's action probabilities, reducing the incentive to make large policy updates even when the advantage estimate is large.

(d) Because clipping removes negative advantages, ensuring that all gradient steps are in a purely improving direction.

Correct option: (c)

6. The PPO "uses a few other tricks to keep new policies close to old". Which mechanism is *directly* responsible for this in PPO-Clip?

(a) The introduction of a KL-divergence penalty term whose coefficient is kept fixed during training.

(b) A projection of the policy parameters onto an ℓ_2 ball around the previous parameters after each gradient step.

(c) The explicit clipping of the probability ratio between new and old policies in the objective, which discourages large changes to the policy even when the advantage estimate is large.

(d) The use of a deterministic policy that eliminates stochasticity in action selection.

Correct option: (c)

7. In the Q-learning update rule, which of the following correctly represents the update of the action-value function for a transition (s, a, r, s')?

(a) $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')]$

(b) $Q(s,a) \leftarrow (1-\alpha)\, Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')]$

(c) $Q(s,a) \leftarrow (1-\gamma)\, Q(s,a) + \gamma[r + \alpha \max_{a'} Q(s',a')]$

(d) $Q(s,a) \leftarrow r + \gamma \max_{a'} Q(s',a')$

Correct option: (b)

8. The DQN as an off-policy temporal-difference method. Which statement best explains why DQN is classified as *off-policy*?

(a) Because it assumes a known model of the environment dynamics and never interacts with the environment.

(b) Because it uses a separate target network and therefore never directly follows its own policy.

(c) Because it learns the value of the greedy policy (via $\max_{a'} Q(s',a')$) even though actions are selected using an ϵ-greedy behaviour policy.

(d) Because it always samples actions uniformly at random, independent of the learned Q-values.

Correct option: (c)

9. In the Gridworld example, the game state is represented as a $4 \times 4 \times 4$ tensor, with each slice encoding the position of a specific object on the board. What is the number of input neurons used by the neural network to consume this state?
   (a) 4
   (b) 12
   (c) 16
   (d) 64

   Correct option: (d)
10. The modified Q-function used in DQN where the network takes a state s and outputs a vector of Q-values, one per action. What is the primary advantage of this vector-valued Q-function compared to the original scalar Q-function $Q(s, a)$ that takes both state and action as input?
   (a) It completely removes the need for a discount factor γ.
   (b) It allows the network to share computation across all actions, requiring only one forward pass per state instead of one forward pass per (s, a) pair.
   (c) It guarantees that the learned policy is always optimal for any environment with a finite number of actions.
   (d) It prevents overfitting by forcing all actions to have identical Q-values for a given state.

   Correct option: (b)
11. The use of *reward clipping* in DQN across different Atari games. What is the main purpose of reward clipping as described there?
   (a) To ensure that rewards monotonically increase over time.
   (b) To scale rewards so that the sum of all rewards in an episode is always equal to 1.
   (c) To normalize the reward magnitudes across different games so that large raw rewards do not dominate learning compared to small raw rewards.
   (d) To prevent the agent from receiving any negative rewards, thereby simplifying the optimization problem.

   Correct option: (c)
12. In the Gridworld DQN description, an ϵ-greedy policy is used for action selection. Which of the following correctly reflects the strategy for choosing actions and controlling ϵ?
   (a) ϵ is fixed at 0.5 for all time; with probability 0.5 a random action is chosen, otherwise the worst predicted action is chosen.
   (b) ϵ starts at a high value (e.g., 1) and is slowly decreased over training; with probability ϵ a random action is chosen, and with probability $1 - \epsilon$ the action with the highest predicted Q-value is chosen.
   (c) ϵ is always set to 0; the agent always chooses the action with the highest predicted Q-value.
   (d) ϵ is set to 1 only at test time to encourage maximum exploration.

   Correct option: (b)
13. In the policy gradient, an episode-based loss of the form

$$\sum_t \log p(y_t \mid x_t; \theta)\, A_t$$

   is used, where A_t is an advantage term derived from rewards. Why is using a reward-based advantage A_t preferable to treating all sampled actions in a (possibly losing) episode equally?
   (a) Because using A_t ensures that only actions from winning episodes are ever updated.
   (b) Because a scalar advantage A_t removes the need to compute gradients with respect to θ.
   (c) Because some actions in a losing episode may still be beneficial; a shaped advantage allows the algorithm to push up probabilities for relatively good actions and push down probabilities for relatively bad actions instead of blaming all actions equally.
   (d) Because A_t guarantees that the policy converges in a single episode regardless of the reward structure.

Correct option: (c)

14. For the Q-values defined, which Bellman equation is consistent with the description "immediate reward of entering s plus the discounted expected utility of performing the best action after taking a in s"?
    (a) $Q(s,a) = R(s) + \gamma \sum_{s'} P(s' \mid s,a)\, V(s')$
    (b) $Q(s,a) = R(s) + \gamma \sum_{s'} P(s' \mid s,a)\, \max_{a'} Q(s',a')$
    (c) $Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s' \mid s')\, \max_{a'} Q(s,a')$
    (d) $Q(s,a) = \gamma \sum_{s'} P(s' \mid s,a)\, Q(s',a)$

    Correct option: (b)

15. In the temporal-difference learning view, after observing a transition $\langle s_1, r_1, a, s_2 \rangle$, which expression best matches the TD error used to update $Q(s_1,a)$?
    (a) $\delta = r_1 - Q(s_1,a)$
    (b) $\delta = \gamma \max_{a'} Q(s_2,a') - Q(s_1,a)$
    (c) $\delta = (R(s_1) + \gamma \max_{a'} Q(s_2,a')) - Q(s_1,a)$
    (d) $\delta = (R(s_2) + \gamma \max_{a'} Q(s_1,a')) - Q(s_2,a)$

    Correct option: (c)

16. The Q-learning update rule can be written in two algebraically equivalent forms. In the notation, which pair of updates is equivalent for a transition $\langle s, r, a, s' \rangle$ (assuming $R(s)$ has just been set to r)?
    (a) $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha r$ and $Q(s,a) \leftarrow Q(s,a) + \alpha(r - Q(s,a))$
    (b) $Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$ and $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a'))$
    (c) $Q(s,a) \leftarrow Q(s,a) + \alpha(\gamma \max_{a'} Q(s',a') - Q(s,a))$ and $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\gamma \max_{a'} Q(s',a')$
    (d) $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha r$ and $Q(s,a) \leftarrow (1-\gamma)Q(s,a) + \gamma r$

    Correct option: (b)

17. Consider the convergence discussion. Which condition on the learning rate α is closest to the one given as sufficient (in spirit) for Q-learning to approach the optimal Q-values?
    (a) α increases linearly with $N(s,a)$ so that new experiences dominate older ones.
    (b) α is kept constant and large for all state–action pairs to speed up convergence.
    (c) α is decreased as $N(s,a)$ increases, for example using a schedule such as $\alpha = \frac{10}{9+N(s,a)}$.
    (d) α is set to zero once a state–action pair has been visited at least once, freezing $Q(s,a)$ thereafter.

    Correct option: (c)

18. The remarks that Q-learning "does not care about the policy that the agent is following." Which of the following is the most precise interpretation of this statement?
    (a) The Q-learning update assumes that the agent always follows a random policy in the future.
    (b) The Q-learning update uses the greedy policy (via $\max_{a'} Q(s',a')$) for its prediction, regardless of whether the behaviour policy actually behaves greedily.
    (c) The Q-learning update explicitly uses the current behaviour policy π to weight future Q-values.
    (d) The Q-learning update only applies when the behaviour policy is optimal.

    Correct option: (b)

19. In the reinforcement learning setting described, which of the following best characterizes the agent–environment interaction?
    (a) The agent is given a fixed dataset of state–action–label triples and never interacts with the environment.
    (b) The agent observes the current state and reward at each time step, then chooses an action, and its goal is to maximize the total *discounted* reward over time.

(c) The agent observes only rewards but never observes states, and its goal is to maximize the undiscounted sum of rewards.

(d) The agent chooses all actions in advance before seeing any states or rewards, and then passively receives a single terminal reward.

Correct option: (b)

20. Which of the following is *not* listed as a key challenge of reinforcement learning?

(a) Rewards are received infrequently, making it hard to determine which actions led to them.

(b) Actions can have long-term effects, so a locally bad action may lead to large future rewards.

(c) The agent must trade off exploration and exploitation over time.

(d) The agent always knows the transition probabilities and reward function exactly.

Correct option: (d)

21. In the *passive* reinforcement learning setting considered, what is the primary objective of the agent?

(a) To learn the optimal policy that maximizes long-term reward over all policies.

(b) To learn the transition probabilities only, assuming the reward function is known.

(c) To follow a fixed policy and learn the expected utility $V(s)$ of each state under that policy.

(d) To generate as many different policies as possible without evaluating them.

Correct option: (c)

22. A simple exploration strategy chooses a random action with probability ε and the best known action with probability $1 - \varepsilon$. Why is it suggested to *decrease* ε over time?

(a) To ensure that the agent always explores at the end of training rather than at the beginning.

(b) Because decreasing ε forces the agent to ignore rewards and focus on transitions.

(c) So that the agent explores more when its knowledge is poor (early on), and exploits more once it has gathered better estimates of utilities.

(d) Because a fixed ε guarantees convergence to a suboptimal policy.

Correct option: (c)

23. In the grid world transition model, an intended move succeeds with probability 0.8, and with probability 0.1 each the action results in a 90° left or right turn. If the robot is in state s_{14} (top-right corner) and chooses the action *right*, what is the probability that it remains in s_{14} after the transition?

(a) 0.1

(b) 0.2

(c) 0.8

(d) 0.9

Correct option: (d)

24. The relationship between the optimal value function $V^*(s)$ and the optimal action-value function $Q^*(s, a)$ for a given state s and action a in the grid world. Which equation is consistent with this definition?

(a) $Q^*(s, a) = R(s) + \gamma \max_{a'} Q^*(s, a')$

(b) $Q^*(s, a) = \sum_{s'} P(s' \mid s, a)\, V^*(s')$

(c) $Q^*(s, a) = R(s) + \sum_{s'} P(s' \mid s, a)\, V^*(s')$

(d) $Q^*(s, a) = V^*(s) - \gamma \sum_{s'} P(s' \mid s, a)$

Correct option: (b)

25. For the grid world with $R(s) = -0.04$ for non-goal states, provides the following optimal state values

$V^*(s)$ (for $\gamma = 1$):

	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	0.705	0.655	0.611	0.388
$i=2$	0.762	X	0.660	-1
$i=3$	0.812	0.868	0.918	1

Using the transition model (intended direction with probability 0.8, 90° left and right with probability 0.1 each, and staying in place when hitting a wall), which action maximizes $Q^*(s,a)$ at state s_{13}?

(a) Up
(b) Down
(c) Left
(d) Right

Correct option: (c)

26. For an MDP with reward function R and transition model P, which of the following equations correctly gives the Bellman *optimality* equation for $V^*(s)$?
    (a) $V^*(s) = R(s) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V^*(s')$
    (b) $V^*(s) = R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a) V^*(s')$
    (c) $V^*(s) = \max_a R(s) + \sum_{s'} P(s' \mid s, a) V^*(s')$
    (d) $V^*(s) = \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) V^*(s')$

    Correct option: (b)
27. In value iteration, let $V_i(s)$ denote the value of state s at iteration i. Which of the following update rules matches the algorithm described?
    (a) $V_{i+1}(s) \leftarrow R(s) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V_i(s')$
    (b) $V_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a) V_{i+1}(s')$
    (c) $V_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a) V_i(s')$
    (d) $V_{i+1}(s) \leftarrow \max(V_i(s), R(s))$

    Correct option: (c)
28. Which termination criterion is consistent with the stopping rule for value iteration?
    (a) Stop when $\sum_s |V_{i+1}(s) - V_i(s)| < \varepsilon$.
    (b) Stop when $\max_s |V_{i+1}(s) - V_i(s)| < \varepsilon$.
    (c) Stop when $\min_s |V_{i+1}(s) - V_i(s)| < \varepsilon$.
    (d) Stop when $V_{i+1}(s) = V_i(s)$ for at least one state s.

    Correct option: (b)
29. In the grid-world example with $R(s) = -0.04$ for non-terminal states and $\gamma = 1$, what qualitative behaviour of $V_i(s)$ is emphasized during value iteration?
    (a) $V_i(s)$ immediately jumps to its optimal value in a single iteration for all states.
    (b) $V_i(s)$ stays at zero for all states until the terminal states are reached in an episode.
    (c) $V_i(s)$ for non-terminal states first becomes negative, then gradually increases as paths to the +1 goal are discovered.
    (d) $V_i(s)$ oscillates randomly between positive and negative values with no clear pattern.

    Correct option: (c)
30. Regarding convergence, which statement best reflects about repeated application of the Bellman update in value iteration?
    (a) If the Bellman update is applied infinitely often to all states, the sequence V_i is guaranteed to converge to V^*.
    (b) Convergence to V^* is only guaranteed if the policy remains fixed during all iterations.
    (c) Convergence to V^* requires that the transition probabilities $P(s' \mid s, a)$ change over time.
    (d) The Bellman update may never converge, even if applied infinitely often, unless rewards are all zero.

Correct option: (a)

31. The contrasts *synchronous* and *asynchronous* value iteration. Which of the following correctly distinguishes them?
    (a) Synchronous value iteration updates one state at a time; asynchronous updates all states simultaneously.
    (b) Synchronous value iteration uses $V_i(s')$ to compute all $V_{i+1}(s)$ in a sweep; asynchronous iteration updates state values one at a time in any order, immediately reusing updated values.
    (c) Synchronous value iteration is model-free; asynchronous value iteration is model-based.
    (d) Synchronous value iteration is only applicable to deterministic MDPs; asynchronous value iteration is only for stochastic MDPs.

    Correct option: (b)

32. After value iteration has converged to V^*, how is an optimal stationary policy π^* recovered?
    (a) $\pi^*(s)$ chooses an action uniformly at random in every state s.
    (b) $\pi^*(s) = \arg\max_a R(s)$ for all s.
    (c) $\pi^*(s) = \arg\max_a \sum_{s'} P(s' \mid s, a)\, V^*(s')$ for all s.
    (d) $\pi^*(s)$ is the action that was last executed in state s during value iteration.

    Correct option: (c)

33. In policy iteration, which pair of equations correctly describes the *policy evaluation* and *policy improvement* steps for a current policy π_i?
    (a) Policy evaluation: $V^{\pi_i}(s) = R(s)$; Policy improvement: $\pi_{i+1}(s) = \arg\max_a R(s)$.
    (b) Policy evaluation: $V^{\pi_i}(s) = R(s) + \gamma \sum_{s'} P(s' \mid s, \pi_i(s)) V^{\pi_i}(s')$; Policy improvement: $\pi_{i+1}(s) = \arg\max_a \sum_{s'} P(s' \mid s, a) V^{\pi_i}(s')$.
    (c) Policy evaluation: $V^{\pi_i}(s) = \max_a \sum_{s'} P(s' \mid s, a) V^{\pi_i}(s')$; Policy improvement: $\pi_{i+1}(s) = \pi_i(s)$.
    (d) Policy evaluation: $V^{\pi_i}(s) = \sum_{s'} P(s' \mid s, \pi_{i+1}(s)) V^{\pi_i}(s')$; Policy improvement: $\pi_{i+1}(s) = \pi_i(s)$.

    Correct option: (b)

34. What condition is used to determine when policy iteration has terminated?
    (a) When $V^{\pi_i}(s) = 0$ for all s.
    (b) When the Bellman optimality equation holds for the current V^{π_i}.
    (c) When $\pi_{i+1} = \pi_i$, i.e., the policy does not change between iterations.
    (d) When the sum of rewards in one episode reaches a predefined threshold.

    Correct option: (c)

35. Policy evaluation can be performed exactly by solving a system of linear equations. If there are n states, what is the time complexity of this exact solution using standard linear algebra methods?
    (a) $O(n)$
    (b) $O(n^2)$
    (c) $O(n^3)$
    (d) $O(2^n)$

    Correct option: (c)

36. In approximate (iterative) policy evaluation, which method is described ?
    (a) Performing gradient descent directly on the Bellman optimality residual.
    (b) Performing repeated simplified Bellman updates using the fixed policy π, similar to value iteration but without the $\max_a$.
    (c) Sampling a single trajectory and setting $V^\pi(s)$ equal to the total return observed from that trajectory.
    (d) Randomly reinitializing $V^\pi(s)$ at every iteration until convergence is observed.

    Correct option: (b)

37. Which of the following correctly compares value iteration with policy iteration?
    (a) Value iteration interleaves partial policy evaluation and improvement at every step; policy iteration

performs full policy evaluation followed by policy improvement.

(b) Value iteration and policy iteration are identical algorithms with different names.

(c) Policy iteration only works when the transition model is unknown; value iteration requires a known model.

(d) Value iteration always converges faster than policy iteration, regardless of the MDP.

Correct option: (a)

38. Suppose you write down the Bellman optimality equation and the policy-evaluation equation *for the same state* s_{11} with $\pi(s_{11}) = \textit{down}$. Which pair correctly identifies which equation includes the $\max_a$ operator?

(a) Both equations include a $\max_a$ over actions.

(b) Only the policy-evaluation equation includes a $\max_a$; the Bellman optimality equation does not.

(c) Only the Bellman optimality equation includes a $\max_a$; the policy-evaluation equation uses the fixed action $\pi(s_{11})$.

(d) Neither equation includes a $\max_a$.

Correct option: (c)

39. Which of the following statements about the relationship between $V^*(s)$ and the $V^{\pi_i}(s)$ encountered during policy iteration is most accurate?

(a) Each $V^{\pi_i}(s)$ is always equal to $V^*(s)$ for all i.

(b) The sequence of policies π_i and their values V^{π_i} is constructed so that, under mild conditions, $V^{\pi_i}(s)$ converges to $V^*(s)$ as i increases.

(c) The first policy's value $V^{\pi_0}(s)$ is always strictly better than $V^*(s)$.

(d) Policy iteration guarantees that $V^{\pi_1}(s)$ equals $V^*(s)$ after a single iteration.

Correct option: (b)

40. Consider applying value iteration to an MDP with n states. At each iteration, a full sweep updates $V_i(s)$ for all states. Ignoring the cost of computing the $\max_a$ and the sum over s', how many value updates are performed after k iterations?

(a) k

(b) n

(c) $k + n$

(d) kn

Correct option: (d)