Reinforcement Learning: A Comprehensive Question Bank

Generated based on the provided text for Taha Majlesi

July 12, 2025

Contents

Part I
Multiple Choice Questions

1 Questions

1. **What is the primary focus of Reinforcement Learning?**

   a. Learning from labeled data.
   b. Discovering hidden structures in data.
   c. Goal-directed learning from interaction.
   d. Predicting continuous values.

   Correct Answer: c

2. **How does an RL agent learn?**

   a. By being given explicit instructions.
   b. Through a process of trial and error.
   c. By analyzing a static dataset.
   d. By finding correlations in unlabeled data.

   Correct Answer: b

3. **What characterizes the problem RL addresses?**

   a. Single decision-making with certainty.
   b. Non-sequential tasks.
   c. Sequential decision-making under uncertainty.
   d. Problems with immediately obvious consequences.

   Correct Answer: c

4. **In the robotic arm example, what is the 'reward'?**

   a. The coordinates (x,y,z).
   b. The robotic arm itself.
   c. A signal like +1 for success or -1 for failure.
   d. The vast, labeled dataset.

   Correct Answer: c

5. **What is a 'policy' in RL?**

   a. The environment's rules.
   b. The agent's strategy for choosing actions.
   c. The cumulative reward.

d. A single action.

Correct Answer: b

6. **Which statement about the data in RL is true?**

   a. It is static and pre-collected.

   b. It is always labeled with the correct action.

   c. It is generated by the agent's own exploration.

   d. It is independent and identically distributed (i.i.d.).

   Correct Answer: c

7. **What is the key difference in feedback between Supervised Learning and RL?**

   a. Supervised feedback is delayed, RL feedback is immediate.

   b. Supervised feedback is evaluative, RL feedback is instructive.

   c. Supervised feedback is instructive, RL feedback is evaluative.

   d. There is no feedback in Supervised Learning.

   Correct Answer: c

8. **What is the main goal of Unsupervised Learning?**

   a. To maximize a reward signal.

   b. To learn a mapping from inputs to outputs.

   c. To discover hidden patterns in unlabeled data.

   d. To learn a sequence of actions.

   Correct Answer: c

9. **The exploration-exploitation dilemma is a direct consequence of what?**

   a. Using a static dataset.

   b. The agent actively generating its own data.

   c. Having instructive feedback.

   d. The problem being fully observable.

   Correct Answer: b

10. **What does AI Planning assume that RL does not?**

    a. The agent must learn from trial and error.

    b. A known model of the environment is available.

    c. The goal is to maximize a reward.

    d. The problem is sequential.

    Correct Answer: b

11. **What is the 'agent' in the agent-environment loop?**

    a. The world in which the task is performed.

    b. The learner and decision-maker.

    c. The reward signal.

    d. The set of all possible states.

    Correct Answer: b

12. **What does the agent receive from the environment at each time step?**

    a. A new policy.

    b. A state observation and a reward.

    c. A set of instructions.

    d. An updated model of the world.

    Correct Answer: b

13. **A sequence of states, actions, and rewards $(S_0, A_0, R_1, S_1, ...)$ is called a:**

    a. Policy.

    b. Model.

    c. Trajectory.

    d. State space.

    Correct Answer: c

14. **What is the primary purpose of a Markov Decision Process (MDP)?**

    a. To provide a conceptual understanding of learning.

    b. To define the agent's policy.

    c. To provide a formal mathematical framework for the RL problem.

    d. To store the agent's experiences.

    Correct Answer: c

15. **What are the components of an MDP tuple?**

    a. (S, A, P, R, γ)

    b. (State, Action, Reward, Policy)

    c. (Agent, Environment, Trajectory, Goal)

    d. (Q-value, V-value, Policy, Reward)

    Correct Answer: a

16. **In an MDP, what does 'P' represent?**

    a. The policy.

b. The reward function.

c. The set of all possible states.

d. The transition probability function.

Correct Answer: d

17. **The Markov Property states that the future is independent of the past given the...**

a. Entire history.

b. Present.

c. Policy.

d. Reward function.

Correct Answer: b

18. **What is the significance of the Markov Property?**

a. It makes the environment deterministic.

b. It ensures rewards are always positive.

c. It allows the policy to be a function of the current state alone.

d. It eliminates the need for a discount factor.

Correct Answer: c

19. **What is a major challenge in applying RL to real-world problems like robotics?**

a. Defining the agent.

b. Constructing a state representation that has the Markov Property.

c. Choosing a discount factor.

d. Deciding on the number of time steps.

Correct Answer: b

20. **What is a POMDP?**

a. A Perfectly Observable Markov Decision Process.

b. A Policy-Optimized Markov Decision Process.

c. A Partially Observable Markov Decision Process.

d. A Probabilistic Optimal Markov Decision Process.

Correct Answer: c

21. **What is the Reward Hypothesis?**

a. All rewards must be positive.

b. All goals can be framed as maximizing a cumulative scalar reward.

c. Rewards are a function of the policy.

d. The agent must receive a reward at every time step.

Correct Answer: b

22. **What is "reward hacking"?**

a. When an agent finds a loophole to maximize reward without achieving the intended goal.

b. When an agent fails to find any rewards.

c. The process of designing a good reward function.

d. When the reward signal is too sparse.

Correct Answer: a

23. **What is the 'return' (G_t)?**

a. The immediate reward at time t.

b. The agent's policy.

c. The cumulative future reward from time step t.

d. The probability of reaching a terminal state.

Correct Answer: c

24. **For which type of task is an undiscounted sum of rewards typically used?**

a. Continuing tasks.

b. Episodic tasks.

c. All tasks.

d. Tasks with no terminal state.

Correct Answer: b

25. **What is the purpose of the discount factor (γ)?**

a. To make the agent prioritize only immediate rewards.

b. To ensure the sum of rewards in continuing tasks remains finite.

c. To increase the learning rate.

d. To define the state space.

Correct Answer: b

26. **A discount factor (γ) close to 0 makes the agent...**

a. Farsighted.

b. Myopic (short-sighted).

c. Value all future rewards equally.

d. Ignore all rewards.

Correct Answer: b

27. **The recursive structure $G_t = R_{t+1} + \gamma G_{t+1}$ is foundational for what?**

   a. The Markov Property.

   b. The Bellman equations.

   c. The agent-environment loop.

   d. The definition of a policy.

   Correct Answer: b

28. **What is a stochastic policy, $\pi(a|s)$?**

   a. A policy that always chooses the same action for a given state.

   b. A policy that provides a probability distribution over actions for a state.

   c. A policy that ignores the state.

   d. A policy that is guaranteed to be optimal.

   Correct Answer: b

29. **What is the state-value function, $V^\pi(s)$?**

   a. The immediate reward in state s.

   b. The probability of being in state s.

   c. The expected return starting from state s and following policy π.

   d. The best possible action to take in state s.

   Correct Answer: c

30. **What is the action-value function, $Q^\pi(s, a)$?**

   a. The value of a state s, regardless of the action.

   b. The expected return after taking action a in state s and then following policy π.

   c. The probability of taking action a in state s.

   d. The immediate reward for taking action a in state s.

   Correct Answer: b

31. **Why is learning the Q-function often preferred for model-free control?**

   a. It requires a model of the environment.

   b. It is simpler to calculate than the V-function.

   c. It directly tells the agent the value of each action, making action selection easy.

   d. It only works for deterministic policies.

   Correct Answer: c

32. **If you have the optimal Q-function, $Q^*(s,a)$, how do you find the optimal action?**

    a. By choosing an action randomly.

    b. By solving a system of linear equations.

    c. By selecting the action ‘a‘ that maximizes $Q^*(s,a)$.

    d. By looking at the V-function.

    Correct Answer: c

33. **What do the Bellman equations provide?**

    a. A way to define the agent’s policy directly.

    b. A recursive framework for computing and learning value functions.

    c. A method for exploring the environment.

    d. A proof of the Reward Hypothesis.

    Correct Answer: b

34. **The Bellman Expectation Equation describes the value function for...**

    a. An optimal policy.

    b. A random policy.

    c. A given, fixed policy π.

    d. A model-based agent.

    Correct Answer: c

35. **What distinguishes the Bellman Optimality Equation from the Expectation Equation?**

    a. The Optimality Equation is linear.

    b. The Optimality Equation includes a maximization (‘max‘) over actions.

    c. The Optimality Equation does not use a discount factor.

    d. The Optimality Equation is only for V-functions, not Q-functions.

    Correct Answer: b

36. **The Principle of Optimality is embodied in which equations?**

    a. The laws of physics.

    b. The agent-environment loop.

    c. The Bellman optimality equations.

    d. The definition of a trajectory.

    Correct Answer: c

37. **Value-based RL algorithms are based on solving which equations?**

a. Maxwell's equations.
b. The Bellman optimality equations.
c. The Policy Gradient Theorem.
d. Linear regression equations.

Correct Answer: b

38. **What is the core idea of Model-Based RL?**

a. To ignore the environment's dynamics.
b. To learn a model of the environment and then use it to plan.
c. To directly learn a policy without a value function.
d. To only learn a value function.

Correct Answer: b

39. **What is the main advantage of Model-Based RL?**

a. Simplicity of implementation.
b. Guaranteed optimal performance.
c. Superior sample efficiency.
d. It works well in continuous action spaces.

Correct Answer: c

40. **What is the main weakness of Model-Based RL?**

a. It is very sample inefficient.
b. Performance is bottlenecked by the accuracy of the learned model.
c. It cannot be used for planning.
d. It only works for episodic tasks.

Correct Answer: b

41. **The Dyna-Q architecture integrates...**

a. Policy gradient and value-based methods.
b. Supervised and unsupervised learning.
c. Direct RL, model learning, and planning.
d. Exploration and exploitation.

Correct Answer: c

42. **What is the core idea of Value-Based RL methods?**

a. To learn a model of the world.
b. To learn a value function and derive a policy from it.
c. To directly parameterize and optimize a policy.

d. To use Monte Carlo Tree Search.

Correct Answer: b

43. **Q-learning is an example of what kind of algorithm?**

a. Model-Based.

b. Policy Gradient.

c. Actor-Critic.

d. Value-Based.

Correct Answer: d

44. **What does it mean for Q-learning to be ”off-policy”?**

a. It learns a policy that is different from the one used to generate data.

b. It does not have a policy.

c. It learns the policy directly.

d. It requires a model of the environment.

Correct Answer: a

45. **What is the ”TD Target” in the Q-learning update rule?**

a. $Q(S_t, A_t)$

b. α

c. $R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a')$

d. The learning rate.

Correct Answer: c

46. **What innovation was critical for making Deep Q-Networks (DQN) stable?**

a. Using a linear function approximator.

b. Using an on-policy learning rule.

c. Using an experience replay buffer.

d. Removing the discount factor.

Correct Answer: c

47. **What is the core idea of Policy Gradient methods?**

a. To learn a Q-function.

b. To learn a model of the environment.

c. To directly learn a parameterized policy.

d. To use the Bellman equations to find a policy.

Correct Answer: c

48. **What is a key advantage of Policy Gradient methods?**

    a. They are always more sample efficient than value-based methods.

    b. They are naturally applicable to continuous action spaces.

    c. They have very low variance.

    d. They are off-policy.

    Correct Answer: b

49. **The Policy Gradient Theorem provides a way to compute the gradient without needing...**

    a. The reward signal.

    b. The agent's policy.

    c. The derivative of the state distribution.

    d. The action-value function.

    Correct Answer: c

50. **What does the REINFORCE algorithm use as an estimate for $Q^{\pi}(S_t, A_t)$?**

    a. The immediate reward R_{t+1}.

    b. The full Monte-Carlo return G_t.

    c. A learned value function $V(S_t)$.

    d. The TD Target.

    Correct Answer: b

51. **What is the primary weakness of the REINFORCE algorithm?**

    a. It is biased.

    b. It only works for deterministic policies.

    c. It has high variance due to the Monte-Carlo return.

    d. It cannot be used with neural networks.

    Correct Answer: c

52. **How do Actor-Critic methods primarily reduce the variance of policy gradients?**

    a. By using a very small learning rate.

    b. By using a learned value function (the Critic) as a baseline.

    c. By using Monte-Carlo returns.

    d. By learning a model of the environment.

    Correct Answer: b

53. **In an Actor-Critic architecture, what is the 'Actor'?**

a. The value function.
b. The policy.
c. The environment model.
d. The reward function.

Correct Answer: b

54. **What is the 'Critic' responsible for?**

a. Selecting actions.
b. Estimating the value function to evaluate the Actor's actions.
c. Storing experiences in a replay buffer.
d. Updating the policy parameters directly.

Correct Answer: b

55. **The Advantage Function $A^\pi(s, a)$ is defined as:**

a. $Q^\pi(s, a) + V^\pi(s)$
b. $V^\pi(s) - Q^\pi(s, a)$
c. $Q^\pi(s, a) - V^\pi(s)$
d. $R_{t+1} + \gamma V^\pi(S_{t+1})$

Correct Answer: c

56. **What key innovation did the A3C algorithm introduce?**

a. Experience Replay.
b. Target Networks.
c. Asynchronous parallel actors to decorrelate data.
d. A model of the environment.

Correct Answer: c

57. **The exploration-exploitation dilemma describes the trade-off between:**

a. Learning a model and learning a policy.
b. Using a value function and a policy function.
c. Choosing the known best action and trying a new action to gain information.
d. Episodic and continuing tasks.

Correct Answer: c

58. **In the ϵ-greedy strategy, what does ϵ represent?**

a. The probability of exploiting the best-known action.
b. The learning rate.
c. The discount factor.

d. The probability of choosing a random action for exploration.

Correct Answer: d

59. **What is the main drawback of ϵ-greedy exploration?**

    a. It is too complex to implement.

    b. It never explores.

    c. Its exploration is undirected and inefficient.

    d. It only works in deterministic environments.

    Correct Answer: c

60. **The UCB (Upper-Confidence-Bound) strategy is based on what principle?**

    a. Pessimism in the face of uncertainty.

    b. Optimism in the face of uncertainty.

    c. Random exploration.

    d. Following an expert.

    Correct Answer: b

61. **Thompson Sampling is a(n) ________ approach to exploration.**

    a. Greedy.

    b. Random.

    c. Bayesian.

    d. Model-based.

    Correct Answer: c

62. **What is the primary role of deep neural networks in Deep Reinforcement Learning (DRL)?**

    a. To store the replay buffer.

    b. To act as powerful function approximators for policies, values, or models.

    c. To define the reward signal.

    d. To ensure the Markov Property holds.

    Correct Answer: b

63. **What does ”end-to-end learning” in DRL refer to?**

    a. Learning from the start of an episode to the end.

    b. Learning to map raw sensory inputs directly to actions.

    c. Learning both a policy and a value function.

    d. Learning a perfect model of the environment.

Correct Answer: b

64. **Which of these was NOT a key innovation of the original DQN paper?**

    a. Deep Q-Network using a CNN.

    b. Experience Replay.

    c. Target Network.

d. Asynchronous actors. **Correct Answer: d**

65. **What problem does the ”target network” in DQN help solve?**

    a. High variance in policy gradients.
    b. The exploration-exploitation dilemma.
    c. Instability from having a rapidly changing TD target.
    d. Inaccurate environment models.

    Correct Answer: c

66. **What is the goal of Imitation Learning (IL)?**

    a. To learn a reward function from an expert.
    b. To learn a task by mimicking an expert demonstrator.
    c. To explore the environment randomly.
    d. To build a model of the world.

    Correct Answer: b

67. **Behavioral Cloning frames imitation learning as a(n) ________ problem.**

    a. Reinforcement Learning.
    b. Unsupervised Learning.
    c. Supervised Learning.
    d. Planning.

    Correct Answer: c

68. **What is the ”covariate shift” problem in Behavioral Cloning?**

    a. The expert provides bad demonstrations.
    b. The agent learns the reward function instead of the policy.
    c. The agent makes a mistake, enters an unseen state, and fails to recover.
    d. The state distribution shifts during training.

    Correct Answer: c

69. **What is the primary goal of Inverse Reinforcement Learning (IRL)?**

    a. To directly mimic the expert’s actions.
    b. To infer the expert’s underlying reward function from their behavior.
    c. To find the optimal policy in a known environment.
    d. To solve the exploration-exploitation dilemma.

Correct Answer: b

70. **What is a key advantage of IRL over Behavioral Cloning?**

    a. It is much simpler to implement.
    b. It is more robust and can generalize better by learning the task's goal.
    c. It does not require expert demonstrations.
    d. It is guaranteed to find the true reward function.

    Correct Answer: b

71. **According to Sutton's "Bitter Lesson," what are the two methods that scale best in AI?**

    a. Human knowledge and logic.
    b. Search and learning.
    c. Supervised and unsupervised learning.
    d. Hardware and software.

    Correct Answer: b

72. **AlphaGo combined which of the following techniques?**

    a. Only Supervised Learning.
    b. Only Reinforcement Learning.
    c. Supervised Learning, Reinforcement Learning, and Monte Carlo Tree Search.
    d. Only Model-Based Planning.

    Correct Answer: c

73. **In AlphaGo, what was the role of the policy network?**

    a. To evaluate the final board position.
    b. To act as the search algorithm.
    c. To narrow the search to promising moves for the MCTS algorithm.
    d. To play the game directly without any search.

    Correct Answer: c

74. **In AlphaGo, what was the role of the value network?**

    a. To suggest which moves to explore.
    b. To evaluate board positions at the end of MCTS simulations.
    c. To store expert games.
    d. To directly choose the winning move.

    Correct Answer: b

75. **Which of the following is a major open challenge in RL?**

    a. Defining what an agent is.
    b. Improving sample efficiency for real-world applications.
    c. Creating a mathematical framework for RL.

d. Applying RL to simple board games.

Correct Answer: b

76. **Model-free methods are often preferred over model-based methods when:**

    a. Real-world samples are very expensive.
    b. A highly accurate model of the environment is easy to learn.
    c. Simulation is cheap and abundant, and asymptotic performance is key.
    d. High sample efficiency is the only priority.

    Correct Answer: c

77. **The term "bootstrapping" in RL refers to:**

    a. Learning from complete episodes only.
    b. Updating estimates based on other learned estimates.
    c. Starting the learning process from scratch.
    d. Using expert demonstrations.

    Correct Answer: b

78. **Which algorithm family is most directly suited for learning stochastic policies?**

    a. Value-Based Methods.
    b. Model-Based Methods.
    c. Policy Gradient Methods.
    d. Imitation Learning.

    Correct Answer: c

79. **What does the "evaluative" nature of RL feedback mean?**

    a. It tells the agent exactly what the correct action was.
    b. It provides no information about the quality of an action.
    c. It indicates how good an action was, but not which action would have been better.
    d. It is always a positive number.

    Correct Answer: c

80. **The entire field of RL is built upon the:**

    a. Policy Gradient Theorem.
    b. Bellman Equations.
    c. Markov Property.
    d. Reward Hypothesis.

    Correct Answer: d

Part II
Explanatory Questions and Answers

2 Questions

1. Explain the core Reinforcement Learning problem and its key characteristics using the robotic arm example.
2. Contrast Reinforcement Learning with Supervised and Unsupervised Learning. Focus on their goals, data requirements, and feedback mechanisms.
3. Describe the agent-environment interaction loop, defining each component (Agent, Environment) and the steps involved in one cycle of interaction (State, Action, Reward).
4. What is a Markov Decision Process (MDP)? List and define its five components.
5. Explain the Markov Property. Why is this property so crucial for simplifying the RL problem?
6. What is the Reward Hypothesis and why is "reward engineering" considered a critical and challenging aspect of applying RL?
7. Differentiate between episodic and continuing tasks. How does the formulation of the return (G_t) change for each, and why is the discount factor (γ) necessary for continuing tasks?
8. Explain the behavioral influence of the discount factor (γ). What does a γ close to 0 versus a γ close to 1 imply about the agent's strategy?
9. Define a policy (π). Distinguish between a stochastic and a deterministic policy.
10. Define the state-value function ($V^{\pi}(s)$) and the action-value function ($Q^{\pi}(s, a)$). What question does each function answer?
11. Explain the practical advantage of learning the action-value function (Q^*) over the state-value function (V^*) for a model-free agent that needs to select actions.
12. What are the Bellman equations? Explain the fundamental difference between the Bellman Expectation Equation and the Bellman Optimality Equation.
13. What is the core principle of Model-Based RL? Describe the two main phases of a typical model-based approach.
14. Discuss the primary strength and the fundamental weakness of Model-Based RL. Why does this create a trade-off with model-free methods?
15. Explain the Q-learning algorithm. Deconstruct its update rule and explain the roles of the learning rate (α), the TD Target, and the TD Error.
16. What does it mean for an algorithm to be "off-policy"? Why was this property of Q-learning essential for the success of Deep Q-Networks (DQN)?
17. What is the main principle behind Policy Gradient methods? What key advantage do they have over value-based methods, particularly in certain types of action spaces?
18. Explain the intuition behind the Policy Gradient Theorem. What do the two main terms, $\nabla_\theta \log \pi_\theta(a|s)$ and $Q^{\pi}(s, a)$, represent?

19. What is the REINFORCE algorithm and what is its major drawback? How does this drawback motivate the development of Actor-Critic methods?

20. Describe the hybrid architecture of Actor-Critic methods. What are the distinct roles of the "Actor" and the "Critic"?

21. What is the Advantage Function, $A^{\pi}(s, a)$? How does it provide a better learning signal for the Actor compared to using the raw return or Q-value?

22. What was the key innovation of the A3C algorithm, and how did it help stabilize the training of deep RL agents?

23. Explain the exploration-exploitation dilemma. Why is balancing the two critical for successful learning?

24. Compare and contrast the exploration strategies of ϵ-greedy, UCB, and Thompson Sampling.

25. What role do deep neural networks play in Deep Reinforcement Learning (DRL)? What is "end-to-end" learning?

26. Describe the three key innovations that made Deep Q-Networks (DQN) successful and stable.

27. What is Imitation Learning? Explain the concept of Behavioral Cloning and its primary limitation, "covariate shift."

28. How does Inverse Reinforcement Learning (IRL) differ from Behavioral Cloning? What is the main advantage of inferring a reward function instead of directly copying actions?

29. Explain Richard Sutton's "Bitter Lesson." How does Reinforcement Learning, and specifically AlphaGo, exemplify this lesson?

30. Describe how AlphaGo integrated multiple AI techniques (Supervised Learning, RL, and Model-Based Search) to achieve its success.

3 Answers

1. **Explain the core Reinforcement Learning problem and its key characteristics using the robotic arm example.**

[redacted]

2. **Contrast Reinforcement Learning with Supervised and Unsupervised Learning. Focus on their goals, data requirements, and feedback mechanisms.**

[redacted]

3. **Describe the agent-environment interaction loop, defining each component (Agent, Environment) and the steps involved in one cycle of interaction (State, Action, Reward).**

[redacted]

4. **What is a Markov Decision Process (MDP)? List and define its five components.**

[redacted]

5. **Explain the Markov Property. Why is this property so crucial for simplifying the RL problem?**

[redacted]

6. **What is the Reward Hypothesis and why is "reward engineering" considered a critical and challenging aspect of applying RL?**

[redacted]

7. **Differentiate between episodic and continuing tasks. How does the formulation of the return (G_t) change for each, and why is the discount factor (γ) necessary for continuing tasks?**

[redacted]

8. **Explain the behavioral influence of the discount factor (γ). What does a γ close to 0 versus a γ close to 1 imply about the agent's strategy?**

[redacted]

9. **Define a policy (π). Distinguish between a stochastic and a deterministic policy.**

[redacted]

10. **Define the state-value function ($V^{\pi}(s)$) and the action-value function ($Q^{\pi}(s, a)$). What question does each function answer?**

[redacted]

11. **Explain the practical advantage of learning the action-value function (Q^{*}) over the state-value function (V^{*}) for a model-free agent that needs to select actions.**

[redacted]

12. **What are the Bellman equations? Explain the fundamental difference between the Bellman Expectation Equation and the Bellman Optimality Equation.**

[redacted]

13. **What is the core principle of Model-Based RL? Describe the two main phases of a typical model-based approach.**

[redacted]

14. **Discuss the primary strength and the fundamental weakness of Model-Based RL. Why does this create a trade-off with model-free methods?**

[redacted]

15. **Explain the Q-learning algorithm. Deconstruct its update rule and explain the roles of the learning rate (α), the TD Target, and the TD Error.**

[redacted]

16. **What does it mean for an algorithm to be "off-policy"? Why was this property of Q-learning essential for the success of Deep Q-Networks (DQN)?**

[redacted]

17. **What is the main principle behind Policy Gradient methods? What key advantage do they have over value-based methods, particularly in certain types of action spaces?**

[redacted]

18. **Explain the intuition behind the Policy Gradient Theorem. What do the two main terms, $\nabla_\theta \log \pi_\theta(a|s)$ and $Q^\pi(s,a)$, represent?**

[redacted]

19. **What is the REINFORCE algorithm and what is its major drawback? How does this drawback motivate the development of Actor-Critic methods?**

[redacted]

20. **Describe the hybrid architecture of Actor-Critic methods. What are the distinct roles of the "Actor" and the "Critic"?**

[redacted]

21. **What is the Advantage Function, $A^{\pi}(s, a)$? How does it provide a better learning signal for the Actor compared to using the raw return or Q-value?**

[redacted]

22. **What was the key innovation of the A3C algorithm, and how did it help stabilize the training of deep RL agents?**

[redacted]

23. **Explain the exploration-exploitation dilemma. Why is balancing the two critical for successful learning?**

[redacted]

24. **Compare and contrast the exploration strategies of ϵ-greedy, UCB, and Thompson Sampling.**

[redacted]

25. **What role do deep neural networks play in Deep Reinforcement Learning (DRL)? What is "end-to-end" learning?**

[redacted]

26. **Describe the three key innovations that made Deep Q-Networks (DQN) successful and stable.**

[redacted]

27. **What is Imitation Learning? Explain the concept of Behavioral Cloning and its primary limitation, ”covariate shift.”**

[redacted]

28. **How does Inverse Reinforcement Learning (IRL) differ from Behavioral Cloning? What is the main advantage of inferring a reward function instead of directly copying actions?**

[redacted]

29. **Explain Richard Sutton’s ”Bitter Lesson.” How does Reinforcement Learn-**

ing, and specifically AlphaGo, exemplify this lesson?

[redacted]

30. **Describe how AlphaGo integrated multiple AI techniques (Supervised Learning, RL, and Model-Based Search) to achieve its success.**

[redacted]