Quiz on Advanced Policy Gradient Methods in Reinforcement Learning

Based on the Analysis by Taha Majlesi

July 17, 2025

Contents

1 Part 1: Multiple-Choice Questions

Instructions: Select the best answer for each of the following questions based on the provided text.

1. What is the primary characteristic of policy-based reinforcement learning methods?

   (A) They learn a value function and derive the policy implicitly.

   (B) They directly parameterize and optimize the agent's policy.

   (C) They require a complete model of the environment's dynamics.

   (D) They can only be used in discrete action spaces.

2. Which family of algorithms learns a state-value function V(s) or an action-value function Q(s,a) to estimate expected future return?

   (A) Policy-based methods

   (B) Model-based methods

   (C) Value-based methods

   (D) Gradient-free methods

3. Actor-Critic methods are described as a hybrid approach because they combine elements of which two other families of methods?

   (A) Model-based and Model-free methods

   (B) Value-based and Policy-based methods

   (C) On-policy and Off-policy methods

   (D) Supervised and Unsupervised learning

4. What is the main advantage of policy-based methods in high-dimensional or continuous action spaces?

   (A) They guarantee finding the global optimum.

   (B) They are simpler to implement than value-based methods.

   (C) Calculating maximum action values can be intractable for value-based methods in these spaces.

   (D) They have inherently lower variance than value-based methods.

5. What is the central objective function, $J(\theta)$, that policy gradient methods aim to maximize?

   (A) The immediate reward at each timestep.

   (B) The probability of the most likely action.

   (C) The expected cumulative reward.

   (D) The entropy of the policy.

6. The term $\nabla_\theta \log \pi_\theta(a|s)$ is known as the:

   (A) Advantage function

(B) Value function

(C) Score function

(D) Reward function

7. According to the intuition behind the REINFORCE algorithm, what happens if a trajectory yields a high total reward?

   (A) The policy parameters are updated to make that trajectory less likely.

   (B) The learning rate is decreased.

   (C) The policy parameters are updated to make that trajectory more likely.

   (D) The value function is reset.

8. What is the most critical flaw of the vanilla policy gradient estimator as used in REINFORCE?

   (A) It is statistically biased.

   (B) It has extremely high variance.

   (C) It only works for deterministic policies.

   (D) It is computationally too expensive.

9. Is the vanilla policy gradient estimator biased or unbiased?

   (A) Biased, because it uses sampled trajectories.

   (B) Unbiased, its expected value is the true gradient.

   (C) Biased, because of the stochastic policy.

   (D) It depends on the learning rate.

10. Which of the following is NOT listed as a source of randomness contributing to high variance in policy gradients?

    (A) Stochastic Policy

    (B) Stochastic Environment Dynamics

    (C) Stochastic Initial State Distribution

    (D) Stochastic Learning Rate

11. What is a direct consequence of high variance in the training process?

    (A) Guaranteed faster convergence.

    (B) The need for a very large learning rate.

    (C) Unstable training and slow convergence.

    (D) Perfect reproducibility across different random seeds.

12. The high variance of the policy gradient is described as a manifestation of what fundamental problem?

    (A) The exploration-exploitation dilemma.

(B) The deadly triad.

(C) Poor credit assignment.

(D) Overfitting to the training data.

13. What is the first and most intuitive technique introduced to improve credit assignment?

    (A) Using a neural network.

    (B) Applying the principle of causality (reward-to-go).

    (C) Using a fixed, large learning rate.

    (D) Adding more layers to the policy network.

14. How does the "reward-to-go" technique modify the learning signal?

    (A) It scales the score function by the total reward of the entire trajectory.

    (B) It scales the score function only by the sum of rewards from the current timestep forward.

    (C) It ignores rewards completely and only uses the score function.

    (D) It uses the average reward of all past episodes.

15. What is the effect of using reward-to-go on the bias of the gradient estimate?

    (A) It introduces a large amount of bias.

    (B) It makes the estimator biased.

    (C) It does not change the expected value (unbiased).

    (D) It removes all bias from the estimate.

16. What is a primary purpose of introducing a discount factor, γ?

    (A) To increase the variance of the gradient estimate.

    (B) To give more weight to rewards that are far in the future.

    (C) To down-weight the influence of uncertain, distant rewards.

    (D) To ensure the sum of rewards is always infinite.

17. What is the core intuition behind using a baseline in policy gradients?

    (A) To increase the absolute magnitude of the returns.

    (B) To learn whether a return was better or worse than expected.

    (C) To make the gradient estimator biased.

    (D) To eliminate the need for a learning rate.

18. What crucial property must a baseline have to avoid introducing bias into the gradient estimate?

    (A) It must be a large positive number.

    (B) It must depend on the action taken.

    (C) It must not depend on the action taken.

(D) It must be a learned neural network.

19. What is the theoretically optimal action-independent baseline that minimizes variance?

    (A) The average reward over the batch.
    (B) The state-value function, $V^\pi(s_t)$.
    (C) The maximum possible reward in the environment.
    (D) Zero.

20. The learning signal $\hat{Q}_{i,t} - V^\pi(s_t)$ is an estimate of what important quantity?

    (A) The TD Error
    (B) The Bellman Error
    (C) The Advantage Function
    (D) The Policy Entropy

21. In an Actor-Critic architecture, what is the role of the "Actor"?

    (A) To estimate the value of states.
    (B) To criticize the actions taken.
    (C) To control the agent's behavior by selecting actions.
    (D) To model the environment's dynamics.

22. What is the role of the "Critic" in an Actor-Critic architecture?

    (A) To select actions.
    (B) To update the policy parameters directly.
    (C) To evaluate the actions taken by the Actor by estimating a value function.
    (D) To store past experiences in a replay buffer.

23. How is the Advantage Function $A^\pi(s_t, a_t)$ formally defined?

    (A) $V^\pi(s_t) - Q^\pi(s_t, a_t)$
    (B) $r(s_t, a_t) + \gamma V^\pi(s_{t+1})$
    (C) $Q^\pi(s_t, a_t) - V^\pi(s_t)$
    (D) $\mathbb{E}[G_t | S_t = s_t]$

24. What does the advantage function measure?

    (A) The total reward of the episode.
    (B) How much better a specific action is compared to the average action in that state.
    (C) The probability of reaching a terminal state.
    (D) The difference between the current policy and the optimal policy.

25. The practical estimate for the advantage function, $r(s_t, a_t) + \gamma \hat{V}^\pi_\phi(s_{t+1}) - \hat{V}^\pi_\phi(s_t)$, is also known as the:

(A) Monte Carlo Error

(B) Score Function

(C) Temporal Difference (TD) Error

(D) Policy Ratio

26. What is the main benefit of using the TD error as the advantage estimate?

(A) It is an unbiased estimate.

(B) It eliminates the need for a Critic.

(C) It allows computing a low-variance advantage estimate using only a single learned state-value function.

(D) It requires training two separate networks for Q and V functions.

27. When training the Critic, what is the key characteristic of a Monte Carlo (MC) evaluation target?

(A) Low variance and biased.

(B) Low variance and unbiased.

(C) High variance and biased.

(D) High variance and unbiased.

28. What is the key characteristic of a Temporal Difference (TD) or bootstrap evaluation target for the Critic?

(A) Biased and low variance.

(B) Unbiased and high variance.

(C) Unbiased and low variance.

(D) Biased and high variance.

29. What is a primary difference between a batch Actor-Critic and an online Actor-Critic algorithm?

(A) The batch method does not use a Critic.

(B) The online method updates after every single step, while the batch method uses a large collection of experiences.

(C) The online method is always more stable.

(D) The batch method cannot use neural networks.

30. What is a potential advantage of using a shared network architecture for the Actor and Critic?

(A) It is always more stable.

(B) It guarantees faster convergence.

(C) It can be more sample-efficient as both components leverage common features.

(D) It simplifies the loss function by removing the value loss.

31. What critical issue, not fully solved by Actor-Critic methods, does Proximal Policy Optimization (PPO) primarily address?

    (A) High bias in the gradient estimate.

    (B) Instability caused by excessively large policy updates.

    (C) The need for a discount factor.

    (D) The credit assignment problem.

32. What is meant by ”policy collapse”?

    (A) When the policy becomes deterministic.

    (B) A catastrophic drop in performance due to a large, destructive policy update.

    (C) When the policy network’s weights become zero.

    (D) When the agent stops exploring.

33. Why are standard on-policy methods like Actor-Critic considered sample-inefficient?

    (A) They require a model of the environment.

    (B) They can only learn from successful episodes.

    (C) They discard data after a single gradient update because it’s no longer representative of the new policy.

    (D) They use too high a learning rate.

34. What was the core idea of Trust Region Policy Optimization (TRPO)?

    (A) To use a very small, fixed learning rate.

    (B) To maximize performance subject to a constraint on how much the policy can change.

    (C) To eliminate the Critic and only use an Actor.

    (D) To use Monte Carlo returns exclusively.

35. How does TRPO measure the change between the old and new policies?

    (A) Mean Squared Error (MSE)

    (B) Kullback-Leibler (KL) divergence

    (C) Cosine Similarity

    (D) Euclidean Distance

36. What is the main drawback of TRPO that motivated the development of PPO?

    (A) It was not effective at preventing policy collapse.

    (B) It was too sample-inefficient.

    (C) Its use of second-order optimization was complex and computationally expensive.

    (D) It could not be used with neural networks.

37. How does PPO capture the benefits of TRPO?

(A) By using the exact same complex optimization.

(B) By using only first-order optimization to approximate a trust region.

(C) By removing the policy update constraint entirely.

(D) By using a much larger batch of data.

38. In PPO, what is the probability ratio $r_t(\theta)$?

(A) The ratio of the new advantage to the old advantage.

(B) The ratio of the new policy's probability of an action to the old policy's probability.

(C) The ratio of the learning rate to the discount factor.

(D) The ratio of the value loss to the policy loss.

39. What is the purpose of the 'clip' function in the PPO-Clip objective?

(A) To ensure the advantage estimate is always positive.

(B) To constrain the policy update ratio $r_t(\theta)$ to a small interval around 1.

(C) To clip the gradients to prevent them from exploding.

(D) To normalize the state inputs.

40. In the PPO-Clip objective, what happens when the advantage $\hat{A}_t$ is positive and the ratio $r_t(\theta)$ exceeds $1 + \epsilon$?

(A) The objective function becomes infinitely large.

(B) The update is skipped for this sample.

(C) There is no further benefit (gradient) from increasing the ratio.

(D) The learning rate is halved.

41. What does the clipping mechanism do when the advantage $\hat{A}_t$ is negative?

(A) It creates a floor, limiting how much the policy can be penalized for a bad action.

(B) It flips the sign of the advantage.

(C) It ignores the sample completely.

(D) It doubles the penalty to discourage the action more strongly.

42. What is the role of the entropy bonus, $S[\pi_\theta](s_t)$, in the full PPO objective?

(A) To make the policy deterministic.

(B) To increase the variance of the updates.

(C) To encourage exploration by keeping the policy stochastic.

(D) To ensure the value function is accurate.

43. What key feature of PPO improves its sample efficiency compared to a standard one-step Actor-Critic?

(A) It uses a much smaller network.

(B) It can perform multiple epochs of minibatch updates on the same batch of data.

(C) It does not require a value function.

(D) It uses a fixed policy throughout training.

44. The text describes PPO's success as a triumph of what?

    (A) Pure mathematical theory.

    (B) Reinforcement learning engineering as much as science.

    (C) Supervised learning techniques.

    (D) Hardware acceleration.

45. Does the PPO clipping mechanism provide a strict mathematical bound on the KL divergence?

    (A) Yes, it is mathematically equivalent to the TRPO constraint.

    (B) No, research has shown it does not provide a strict bound.

    (C) Yes, but only when ϵ is very small.

    (D) The text does not mention this relationship.

46. What is the first step in the evolutionary path of policy gradients described in the synthesis?

    (A) PPO

    (B) Actor-Critic

    (C) REINFORCE

    (D) TRPO

47. The introduction of causality, discounting, and baselines primarily addressed which problem?

    (A) Update instability

    (B) Poor credit assignment and high variance

    (C) The exploration-exploitation dilemma

    (D) Computational complexity

48. The Actor-Critic framework is presented as a solution that uses a learned function for what purpose?

    (A) To model the environment.

    (B) To act as a sophisticated, state-dependent baseline (the Critic).

    (C) To replace the policy with a deterministic mapping.

    (D) To select the learning rate automatically.

49. PPO addresses the ”final piece of the puzzle,” which is identified as:

    (A) High variance

    (B) High bias

    (C) Preventing destructive, large policy updates

(D) Ensuring the value function is optimal

50. Which algorithm is described as a ”default, go-to algorithm” and a benchmark for modern RL?

    (A) DQN

    (B) REINFORCE

    (C) TRPO

    (D) PPO

51. Which of the following is NOT listed as an active area of future research?

    (A) Greater sample efficiency.

    (B) More sophisticated exploration strategies.

    (C) Moving away from neural networks back to linear models.

    (D) Algorithms more robust to hyperparameter choices.

52. The policy $\pi_\theta(a|s)$ is a:

    (A) Single action for a given state.

    (B) Probability distribution over actions given a state.

    (C) Value representing the quality of a state.

    (D) Model of the next state.

53. In value-based methods, how is the policy typically derived?

    (A) It is learned directly by a separate network.

    (B) It is derived implicitly, for example, by taking the action with the highest Q-value.

    (C) It is a uniform random policy.

    (D) It is provided by a human expert.

54. The REINFORCE algorithm updates the policy based on:

    (A) The immediate reward only.

    (B) The total reward of the entire trajectory.

    (C) The TD error.

    (D) The KL divergence.

55. High variance in gradient estimates can lead to:

    (A) A need for a very small learning rate.

    (B) Stable and fast training.

    (C) Consistent results across different random seeds.

    (D) The policy converging to the global optimum.

56. Using the reward-to-go instead of the full return is an application of the principle of:

(A) Causality

(B) Entropy

(C) Duality

(D) Linearity

57. Subtracting a baseline $b(s_t)$ from the return G_t helps to:

(A) Increase the magnitude of the gradient.

(B) Center the learning signal around zero.

(C) Introduce bias to speed up learning.

(D) Remove the need for a discount factor.

58. The transition from REINFORCE with a baseline to Actor-Critic is motivated by using what as the baseline?

(A) A constant value of 1.

(B) The reward from the previous step.

(C) A learned estimate of the state-value function, $\hat{V}^{\pi}(s)$.

(D) The entropy of the policy.

59. In the Actor-Critic paradigm, which component learns a policy?

(A) The Critic

(B) The Actor

(C) The Environment

(D) The Replay Buffer

60. The TD target for training a Critic is $r + \gamma \hat{V}(s')$. This method is called:

(A) Monte Carlo

(B) Bootstrapping

(C) Gradient Descent

(D) Random Search

61. What is a major risk associated with a learning rate that is too large in policy gradient methods?

(A) Very slow convergence.

(B) The policy becoming too stochastic.

(C) Catastrophic performance collapse.

(D) The value function becoming inaccurate.

62. PPO is a simplification of which earlier, more complex algorithm?

(A) DQN

(B) A2C

(C) TRPO

(D) REINFORCE

63. If the PPO probability ratio $r_t(\theta) = 0.8$ and the advantage $\hat{A}_t = -10$, what is the unclipped objective term?

(A) -8

(B) 8

(C) -12.5

(D) 12.5

64. If $\epsilon = 0.2$, the PPO clipping mechanism constrains the probability ratio $r_t(\theta)$ to which range?

0.2, 1.2

0.8, 1.2

-0.2, 0.2

0.0, 1.0

65. The overall PPO objective function combines the policy loss, the value function loss, and what other term?

(A) A KL penalty

(B) An entropy bonus

(C) A regularization term

(D) A curiosity module

66. The core idea of a "trust region" is to:

(A) Trust the value function estimates completely.

(B) Ensure the new policy does not deviate too much from the old one.

(C) Only update the policy in regions of the state space that have been visited frequently.

(D) Use a very high discount factor.

67. What is the main trade-off discussed in the context of training the Critic?

(A) Speed vs. Memory

(B) Exploration vs. Exploitation

(C) Bias vs. Variance

(D) Online vs. Offline

68. A "greedy policy" derived from a Q-function means:

(A) Selecting actions randomly.

(B) Always selecting the action with the highest estimated Q-value.

(C) Selecting the action with the lowest Q-value to encourage exploration.

(D) Following a pre-defined policy.

69. The term ”credit assignment” refers to:

(A) Assigning memory to the neural network.

(B) Determining which actions are responsible for a given outcome (reward).

(C) The process of initializing the policy parameters.

(D) The financial cost of training a model.

70. Why is it inefficient to reinforce every action in a long, successful trajectory equally?

(A) It is computationally too slow.

(B) It incorrectly gives credit to mediocre or bad actions that happened to be in a good trajectory.

(C) It violates the Markov property.

(D) It causes the learning rate to decay too quickly.

71. What does an online Actor-Critic algorithm do?

(A) It only learns when connected to the internet.

(B) It updates the Actor and Critic after every single step (or a small number of steps).

(C) It waits for a full batch of episodes to complete before any updates.

(D) It uses a pre-trained Critic and only updates the Actor.

72. What is a potential downside of a shared network for the Actor and Critic?

(A) It is less sample-efficient.

(B) It requires two separate optimizers.

(C) The loss signals for policy and value must be carefully balanced.

(D) It cannot be used with PPO.

73. The progression from REINFORCE to PPO shows a deliberate strategy of accepting a small amount of bias in exchange for what?

(A) Faster computation per step.

(B) A substantial reduction in variance.

(C) The ability to use smaller networks.

(D) Guaranteed convergence to the global optimum.

74. What does the term ”model-free” signify in reinforcement learning?

(A) The algorithm does not use a neural network model.

(B) The algorithm learns without building an explicit model of the environment’s dynamics.

(C) The algorithm is free of hyperparameters.

(D) The algorithm does not have a policy model.

75. The policy gradient theorem provides a way to compute the gradient of the objective function with respect to what?

    (A) The state values.
    (B) The action values.
    (C) The policy parameters, θ.
    (D) The environment dynamics.

76. What is a major contributor to the ”reproducibility crisis” in RL mentioned in the text?

    (A) Lack of open-source code.
    (B) High variance in gradient estimates.
    (C) Use of different programming languages.
    (D) Insufficient computational power.

77. The use of Generalized Advantage Estimation (GAE) is mentioned as a strong implementation practice for which algorithm?

    (A) DQN
    (B) REINFORCE
    (C) PPO
    (D) Q-learning

2 Part 2: Explanable Questions

Instructions: Provide a detailed explanation for each of the following questions based on the provided text.

1. **Question:** Explain the fundamental difference between value-based and policy-based reinforcement learning methods. Why are policy-based methods particularly well-suited for continuous action spaces?

2. **Question:** What is the Policy Gradient Theorem, and what is the core intuition behind the REINFORCE algorithm's update rule?

3. **Question:** Describe the problem of high variance in policy gradient methods. What are the three main sources of randomness that contribute to it, and what are its negative consequences for the learning process?

4. **Question:** How does the "credit assignment problem" relate to the high variance in the vanilla policy gradient estimator?

5. **Question:** Explain the principle of "causality" in the context of variance reduction and how the "reward-to-go" formulation addresses it.

6. **Question:** What is a baseline in policy gradient methods, and what is its purpose? Explain why a baseline that does not depend on the action can be subtracted from the return without introducing bias.

7. **Question:** What is the theoretically optimal action-independent baseline, and how does this concept provide a logical bridge to the Actor-Critic framework?

8. **Question:** Describe the roles of the "Actor" and the "Critic" in an Actor-Critic architecture. How do they work together?

9. **Question:** Define the Advantage Function, $A^{\pi}(s, a)$. Why is it considered a superior learning signal compared to the raw reward-to-go?

10. **Question:** How can the Advantage Function be estimated practically using only a single learned state-value function, $\hat{V}_{\phi}^{\pi}(s)$? What is this estimate called?

11. **Question:** Compare and contrast the Monte Carlo (MC) and Temporal Difference (TD) methods for training the Critic. What is the fundamental bias-variance trade-off between them?

12. **Question:** What is the primary motivation behind trust region methods like TRPO and PPO? What is the "peril of large policy updates"?

13. **Question:** Explain the core idea of Trust Region Policy Optimization (TRPO) and its main practical drawback that led to the development of PPO.

14. **Question:** Describe the PPO-Clip algorithm. How does the clipped surrogate objective function work to prevent destructive policy updates? Explain its behavior for both positive and negative advantages.

15. **Question:** What is the probability ratio $r_t(\theta)$ in PPO, and what is the role of the hyperparameter ϵ?

16. **Question:** Why is PPO considered more sample-efficient than many standard on-policy Actor-Critic methods?

17. **Question:** The text describes PPO as a ”triumph of reinforcement learning engineering as much as science.” What does this mean?

18. **Question:** Summarize the evolutionary path of policy gradient methods from REINFORCE to PPO, highlighting the specific problem that each major advancement (baselines, Actor-Critic, PPO) was designed to solve.

19. **Question:** What is the difference between a batch and an online Actor-Critic algorithm?

20. **Question:** Explain the architectural choice between using two separate networks versus a shared network for the Actor and Critic, including the pros and cons of each.

21. **Question:** What is the purpose of the entropy bonus term often included in the objective function of algorithms like PPO?

22. **Question:** Why is data from past policies typically discarded in on-policy learning? How does PPO's stable update mechanism partially alleviate this issue?

23. **Question:** Explain the intuition of ”Good stuff is made more likely” and ”Bad stuff is made less likely” in the context of the basic policy gradient update.

24. **Question:** What is ”policy collapse,” and which algorithm is specifically designed to prevent it by constraining policy updates?

25. **Question:** How does the introduction of a discount factor γ help reduce variance?

26. **Question:** Why is it computationally efficient to use the TD error as an advantage estimate instead of learning two separate networks for the Q-function and V-function?

27. **Question:** What is the fundamental trade-off that is accepted when moving from REINFORCE with a baseline to an Actor-Critic method?

28. **Question:** How does PPO's clipped objective create a ”pessimistic bound” on the policy improvement?

29. **Question:** What are some of the active areas of future research in policy gradient methods mentioned in the conclusion?

30. **Question:** Explain why high variance is a major contributor to the ”reproducibility crisis” in reinforcement learning.

3 Answer Key: Multiple-Choice Questions

Q	Ans	Q	Ans	Q	Ans	Q	Ans
1	B	21	C	41	C	61	B
2	C	22	C	42	A	62	C
3	B	23	C	43	C	63	A
4	C	24	B	44	B	64	B
5	C	25	C	45	B	65	B
6	C	26	C	46	B	66	C
7	C	27	D	47	C	67	B
8	B	28	A	48	B	68	B
9	B	29	B	49	D	69	C
10	D	30	C	50	D	70	B
11	C	31	B	51	C	71	B
12	C	32	B	52	B	72	B
13	B	33	C	53	B	73	C
14	B	34	B	54	A	74	B
15	C	35	C	55	A	75	B
16	C	36	B	56	B	76	C
17	B	37	B	57	C	77	B
18	C	38	B	58	B	78	B
19	B	39	B	59	C	79	B
20	C	40	B	60	C	80	C

4 Answer Key: Explanable Questions

1. **Answer: Value-based methods**, like DQN, focus on learning a value function (e.g., $Q(s, a)$) that estimates the expected return of taking an action in a state. The policy is then derived *implicitly* from this value function, typically by choosing the action with the highest value (a greedy policy). **Policy-based methods**, in contrast, *directly* learn a parameterized policy, $\pi_\theta(a|s)$, which is a probability distribution over actions. The parameters θ are optimized to maximize the expected total reward. Policy-based methods are well-suited for **continuous action spaces** because they output a probability distribution (e.g., a Gaussian with a mean and standard deviation) from which an action can be sampled. For value-based methods, finding the action with the maximum Q-value in a continuous space would require an optimization procedure at every single timestep, which is often intractable.

2. **Answer:** The **Policy Gradient Theorem** provides a way to compute the gradient of the expected total reward $J(\theta)$ with respect to the policy parameters θ, allowing for optimization via gradient ascent. The core intuition of the **REINFORCE** update rule is "trial and error." The agent executes a trajectory and calculates the total reward. The update rule then scales the score function, $\nabla_\theta \log \pi_\theta(a|s)$, by this total reward. This means:

   - If the total reward was high (a good outcome), the probabilities of the actions taken in that trajectory are increased ("Good stuff is made more likely").
   - If the total reward was low (a bad outcome), the probabilities of those actions are decreased ("Bad stuff is made less likely").

3. **Answer:** The problem of **high variance** means that the gradient estimates calculated from a batch of trajectories can fluctuate wildly from one batch to the next, making the learning signal very noisy. The three main sources of randomness are:

   (a) **Stochastic Policy:** The policy itself is a probability distribution, so actions can differ even in the same state.

   (b) **Stochastic Environment Dynamics:** The environment's response to an action can be probabilistic.

   (c) **Stochastic Initial State Distribution:** Episodes can start in different states.

   The negative consequences include: **unstable training**, **slow convergence** (requiring a tiny learning rate), **poor reproducibility**, and the **risk of policy collapse**.

4. **Answer:** The "credit assignment problem" refers to the challenge of determining which specific actions in a sequence are responsible for the final outcome. The vanilla policy gradient estimator (REINFORCE) exhibits poor credit assignment because it uses the total reward of the *entire trajectory* to reinforce *every single action* within it. This means that even a mediocre or bad action will be positively reinforced if the overall trajectory outcome was good. This misattribution of credit to all actions, rather than just the ones that were truly beneficial, introduces significant noise and is a primary reason for the estimator's high variance.

5. **Answer:** The principle of **causality** states that an action taken at timestep t can only influence rewards that occur at or after timestep t; it cannot affect past rewards. The vanilla policy gradient violates this by crediting an action with all rewards in the trajectory, including those that came before it. The **"reward-to-go"** formulation addresses this by modifying the

learning signal. Instead of using the total trajectory reward, it scales the score function for an action a_t only by the sum of rewards from timestep t to the end of the episode ($\sum_{t'=t}^{T} r_{t'}$). This removes the noise from causally irrelevant past rewards, reducing variance without introducing bias.

6. **Answer:** A **baseline** is a value, b, that is subtracted from the return term in the policy gradient update. Its purpose is to **reduce variance**. The intuition is that it's more important to know if an outcome was better or worse than expected, rather than its absolute value. Subtracting a baseline centers the learning signal around zero. A baseline b that does not depend on the action can be subtracted without introducing bias because the expected value of the subtracted term, $\mathbb{E}[\nabla_\theta \log p_\theta(\tau) b]$, is zero. This is because the baseline can be pulled out of the gradient calculation, leading to the gradient of a constant ($\nabla_\theta \int p_\theta(\tau) d\tau = \nabla_\theta 1 = 0$), which is zero.

7. **Answer:** The theoretically optimal action-independent baseline is the one that minimizes the variance of the learning signal. This is achieved by subtracting the mean of the random variable. In this case, the random variable is the reward-to-go, and its expected value conditioned on the state is, by definition, the **state-value function,** $V^\pi(s_t)$. This provides a logical bridge to the **Actor-Critic framework** because to use this optimal baseline, the agent needs a way to *estimate* the state-value function. This motivates the introduction of a dedicated function approximator for this purpose, which is precisely the role of the "Critic."

8. **Answer:** In an Actor-Critic architecture, there is a clear division of labor:

   - **The Actor** is the policy, $\pi_\theta(a|s)$. Its role is to control the agent's behavior by selecting actions and to update its parameters to improve its strategy. It is responsible for the "acting."
   - **The Critic** is the value function approximator, $\hat{V}_\phi^\pi(s)$. Its role is to "criticize" or evaluate the actions taken by the Actor. It does not select actions but learns to predict the expected return from states, providing a high-quality, low-variance learning signal (the advantage) to guide the Actor's updates.

   They work together: the Actor acts, the Critic evaluates, and the Actor uses the Critic's evaluation to update its policy.

9. **Answer:** The Advantage Function is formally defined as $A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$. It is considered a superior learning signal because it measures how much better a specific action a_t is compared to the *average* action that would be taken in state s_t. By using the state value $V^\pi(s_t)$ as a baseline, it centers the learning signal. Actions that are merely "average" will have an advantage near zero and won't cause a significant update, while only actions that are substantially better or worse than average will drive learning. This significantly reduces variance compared to using the raw reward-to-go.

10. **Answer:** The Advantage Function can be estimated practically using a single state-value function, $\hat{V}_\phi^\pi(s)$, by leveraging the relationship $Q^\pi(s_t, a_t) \approx r(s_t, a_t) + \gamma V^\pi(s_{t+1})$. By substituting this into the advantage definition, we get: $\hat{A}^\pi(s_t, a_t) \approx r(s_t, a_t) + \gamma \hat{V}_\phi^\pi(s_{t+1}) - \hat{V}_\phi^\pi(s_t)$. This estimate is called the **Temporal Difference (TD) Error**. This is highly efficient as it avoids the need to train a second network for the Q-function.

11. Answer: For training the Critic, the two methods represent a fundamental bias-variance trade-off:

- **Monte Carlo (MC) Method:** Uses the full, empirical discounted return from a state until the end of the episode as the training target. This target is **unbiased** (it's a true sample of the return) but suffers from **high variance** because it depends on a long sequence of random actions and state transitions.
- **Temporal Difference (TD) Method:** Uses a "bootstrapped" target: the immediate reward plus the discounted value of the next state, as estimated by the Critic itself $(r + \gamma \hat{V}_\phi(s'))$. This target is **biased** because it depends on the current, likely imperfect, value estimate. However, it has significantly **lower variance** as it only depends on one step of randomness.

12. Answer: The primary motivation is to solve the problem of **instability caused by excessively large policy updates**. The "peril of large policy updates" refers to the fact that if the learning rate is too large, a single gradient step based on a noisy batch of data can drastically change the policy for the worse. This can lead to a catastrophic drop in performance, known as **policy collapse**, from which the agent may never recover. Trust region methods aim to take the largest possible improvement step without risking this kind of instability.

13. Answer: The core idea of **TRPO** is to maximize the policy's performance objective while satisfying a *constraint* that the new policy cannot deviate too far from the old one. This "trust region" is enforced by limiting the Kullback-Leibler (KL) divergence between the old and new policies. This allows for large, stable update steps. TRPO's main practical **drawback** is its high complexity and computational cost. Enforcing the KL constraint requires a complex, second-order optimization procedure that is impractical for the large neural networks used in modern deep RL.

14. Answer: The **PPO-Clip** algorithm replaces TRPO's hard KL constraint with a simpler, clipped surrogate objective function. This function limits how much the probability ratio $r_t(\theta)$ can change the objective.

- **When Advantage $\hat{A}_t > 0$ (good action):** The objective is $\min(r_t(\theta)\hat{A}_t, (1+\epsilon)\hat{A}_t)$. This creates a "ceiling." Once the policy ratio increases beyond $1+\epsilon$, there is no further incentive to increase it, preventing an overly large update from a single good action.
- **When Advantage $\hat{A}_t < 0$ (bad action):** The objective is $\max(r_t(\theta)\hat{A}_t, (1-\epsilon)\hat{A}_t)$. This creates a "floor." It limits how much the objective can be penalized, preventing a single bad action from excessively damaging the policy.

This clipping mechanism effectively creates a soft constraint that discourages large policy updates.

15. Answer: The **probability ratio** $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ measures how the probability of taking a specific action a_t in state s_t has changed between the new policy (π_θ) and the old policy $(\pi_{\theta_{old}})$ that was used to collect the data. The hyperparameter ϵ defines the clipping range. It determines how far the new policy is allowed to deviate from the old one. For example, if $\epsilon = 0.2$, the ratio $r_t(\theta)$ is effectively constrained to the interval $[0.8, 1.2]$ within the objective function, thus controlling the size of the policy update.

16. **Answer:** PPO is considered more sample-efficient because its stable update mechanism allows it to perform **multiple epochs of minibatch updates on the same batch of collected data**. Standard on-policy methods typically perform only one gradient update and then discard the data because the policy has changed, making the old data off-policy and potentially harmful to use. Because PPO's updates are constrained and less likely to be destructive, it can reuse the same data for several optimization steps, learning more from each batch of experience and thus improving its sample efficiency.

17. **Answer:** This statement means that PPO's success is not just due to a pure, elegant mathematical theory, but also due to its pragmatic design and effective combination of several practical ideas. While motivated by the strong theory of TRPO, PPO itself is a simpler, first-order approximation. Research has shown its clipping mechanism is more of a robust heuristic than a strict mathematical constraint. Its success comes from combining this clever, simple objective with other strong implementation practices (like GAE), making it a highly effective and practical solution—a triumph of smart engineering choices as much as theoretical advancement.

18. **Answer:** The evolutionary path is as follows:

    (a) **REINFORCE:** The starting point, which established the principle of direct policy optimization but suffered from **high variance**.

    (b) **Baselines/Reward-to-Go:** The first advancements, which aimed to solve the **high variance** problem by improving credit assignment. This culminated in the idea of using the state-value function as the optimal baseline.

    (c) **Actor-Critic:** This framework formalized the use of a learned baseline. The Critic learns the value function to provide a low-variance advantage estimate, further tackling the **variance and credit assignment** problems.

    (d) **PPO:** While Actor-Critic methods solve for variance, they don't prevent unstable updates. PPO addresses this final problem of **update instability** by introducing a clipped objective to constrain policy changes, preventing policy collapse.

19. **Answer:** The main difference is the update frequency.

    - **Batch Actor-Critic:** This method first collects a large batch of experiences (e.g., many complete episodes). Then, in a separate update phase, it uses this entire batch to train the Critic and then update the Actor.
    - **Online Actor-Critic:** This method updates both the Actor and Critic much more frequently, typically after every single timestep or a very small number of steps. It allows for faster adaptation but can have higher variance in its updates.

20. **Answer:** The choice is between:

    - **Two Separate Networks:** The Actor and Critic are independent neural networks.
        - **Pros:** Simple, and can be more stable as the gradient updates for one component do not directly interfere with the other's parameters.
        - **Cons:** May be less data-efficient as the networks cannot share learned features about the state.

- **Shared Network:** A single network body extracts features from the state, which then feed into two separate heads for the policy (Actor) and value (Critic).
  - **Pros:** Can be more sample-efficient as both components leverage a common, powerful feature representation.
  - **Cons:** Can introduce more complex training dynamics, as the loss signals for both policy and value must be carefully balanced to avoid interference.

21. Answer: The purpose of the entropy bonus is to **encourage exploration**. Entropy is a measure of randomness or uncertainty in a probability distribution. By adding the policy's entropy to the objective function, the algorithm is incentivized to keep the policy more stochastic (i.e., less deterministic). This prevents the policy from collapsing to a suboptimal deterministic strategy too early and helps the agent to continue exploring its action space, which can lead to finding better solutions.

22. Answer: In on-policy learning, the data used for updates must be collected with the current policy. After a gradient update, the policy changes. The old data is now "off-policy" because it was generated by a different policy. Using this old data would introduce a harmful bias into the gradient estimate, as the experiences do not accurately reflect the behavior of the new policy. PPO's stable update mechanism, which constrains the policy change to be small, means the new policy is not drastically different from the old one. This makes the old data "less" off-policy, allowing it to be reused for multiple update steps without introducing a crippling amount of bias, thus improving sample efficiency.

23. Answer: This intuition describes the core mechanism of the policy gradient update. The update is proportional to the score function (which points in the direction to make an action more likely) multiplied by the total reward.

- **"Good stuff is made more likely":** If a trajectory results in a high, positive reward ("good stuff"), this reward term is positive. The policy parameters are moved in the direction of the score function, increasing the probability of the actions that were taken.
- **"Bad stuff is made less likely":** If a trajectory results in a low or negative reward ("bad stuff"), this reward term is negative. The policy parameters are moved in the opposite direction of the score function, decreasing the probability of the actions that led to the poor outcome.

24. Answer: Policy collapse is a catastrophic drop in performance that occurs when a single, large policy update moves the policy parameters to a very poor region of the optimization landscape, from which it may be difficult or impossible to recover. **Proximal Policy Optimization (PPO)** is the algorithm specifically designed to prevent this by using its clipped surrogate objective to constrain the size of the policy update at each step.

25. Answer: Introducing a discount factor $\gamma < 1$ reduces variance by **down-weighting the influence of highly uncertain, distant rewards**. Rewards that are far in the future are subject to more steps of stochasticity from both the policy and the environment, making them much more variable and less directly attributable to an early action. By reducing their impact on the total return calculation, the discount factor makes the learning signal (the discounted return) less noisy and thus reduces its variance.

26. Answer: It is computationally efficient because it only requires training and maintaining **one neural network (the Critic for the V-function)** instead of two. A naive approach might train one network for $Q(s, a)$ and another for $V(s)$ to compute the advantage $Q(s, a) - V(s)$. This is expensive. By using the TD error, $r + \gamma\hat{V}(s') - \hat{V}(s)$, as the advantage estimate, we can derive this powerful, low-variance learning signal from just the single, learned state-value function, greatly simplifying the architecture and training process.

27. Answer: The fundamental trade-off is accepting a small amount of **bias** in exchange for a substantial reduction in **variance**. The learning signal in REINFORCE with a baseline (using empirical returns) is unbiased but has high variance. In an Actor-Critic method, the learning signal (the advantage estimate) is derived from a learned function approximator (the Critic). Because the Critic is an imperfect estimate of the true value function, it introduces a small amount of bias into the advantage estimate. However, this estimate is far more stable and has much lower variance than the empirical Monte Carlo return, leading to more stable and efficient learning overall.

28. Answer: PPO's clipped objective creates a pessimistic bound by taking the **minimum** of two terms: the normal, unclipped objective ($r_t(\theta)\hat{A}_t$) and the clipped objective ($\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t$). By taking the minimum, the algorithm ensures that the update will not be overly optimistic. If the unclipped objective suggests a very large improvement (e.g., when r_t is large and advantage is positive), the clipped term will be smaller, and the algorithm will use that smaller, more conservative value. This prevents the agent from taking a large, risky step based on a potentially noisy advantage estimate.

29. Answer: The conclusion mentions several active areas of future research:

- Developing algorithms with even greater **sample efficiency**.
- Designing more sophisticated **exploration strategies** for sparse reward tasks.
- Creating algorithms that are more **robust to hyperparameter choices**.
- Proposing direct alternatives to PPO's clipping mechanism that might offer more **principled KL-divergence control** without sacrificing simplicity.

30. Answer: High variance is a major contributor to the reproducibility crisis because it means that the outcome of the training process is highly sensitive to random factors. With high variance, different random seeds—which affect initial network weights, action sampling, and environment responses—can lead to vastly different training outcomes. One run might succeed and learn a strong policy, while another run with a different seed might fail completely (”outlier runs”). This makes it difficult to reliably reproduce reported results and to determine whether a new algorithm is genuinely better or if its success was due to a lucky random seed.