Your grade: 87.50%

Your latest: **87.50%** • Your highest: **87.50%**
To pass you need at least 80%. We keep your highest score.

Next item →

1. Which of the following is true about policy gradient methods? (**Select all that apply**)

1 point

- [x] Policy gradient methods use generalized policy iteration to learn policies directly.

⊗ **This should not be selected**
Incorrect. Value-based methods use generalized policy iteration to learn approximate action values, and indirectly infer a good policy. Policy gradient methods maximize the policy objective to learn policies directly.

- [x] The policy gradient theorem provides a form for the policy gradient that does not contain the gradient of the state distribution μ, which is hard to estimate.

Correct
Correct.

- [x] Policy gradient methods do gradient ascent on the policy objective.

Correct
Correct. Policy gradient methods maximize the policy objective, and hence perform gradient ascent.

- [x] If we have access to the true value function v_π, we can perform unbiased stochastic gradient updates using the result from the Policy Gradient Theorem.

Correct

Correct. We derived this stochastic update by multiplying and dividing by $\pi(A|S)$.

2. Which of the following statements about parameterized policies are true? (**Select all that apply**)

1 / 1 point

- [x] The probability of selecting any action must be greater than or equal to zero.

Correct
Correct! This is one of the conditions for a valid probability distribution.

- [x] For each state, the sum of all the action probabilities must equal to one.

Correct
Correct! This condition is necessary for the function to be a valid probability distribution.

- [ ] The policy must be approximated using linear function approximation.

- [ ] The function used for representing the policy must be a softmax function.

3. Assume you're given the following preferences $h_1 = 44$, $h_2 = 42$, and $h_3 = 38$, corresponding to three different actions (a_1, a_2, a_3), respectively. Under a softmax policy, what is the probability of choosing a_2, rounded to three decimal numbers?

1 / 1 point

- ( ) 0.42
- ( ) 0.879
- ( ) 0.002

- [x] 0.119

Correct
Correct!

4. Which of the following is true about softmax policy? (Select all that apply) 0.5 / 1 point

- [ ] It is used to represent a policy in discrete action spaces.

- [x] It can be parameterized by any function approximator as long as it can output scalar values for each available action, to form a softmax policy.

Correct
Correct. It can use any function approximation from deep artificial neural networks to simple linear features.

- [x] Similar to epsilon-greedy policy, softmax policy cannot approach a deterministic policy.

This should not be selected
Incorrect. Epsilon-greedy policy will always have epsilon probability of selecting a random action but softmax policy can approach a deterministic policy as the preference of one action dominates other preferences.

- [ ] It cannot represent an optimal policy that is stochastic, because it reaches a deterministic policy as one action preference dominates others.

5. What are the differences between using softmax policy over action-values and using softmax policy over action-preferences? (**Select all that apply**) 1 / 1 point

- [x] When using softmax policy over action-values, even if the optimal policy is deterministic, the policy may never approach a deterministic policy.

Correct

Correct. The policy will always select proportional to exponentiated action-values.

- [ ] When using softmax policy over action-values, assuming a tabular representation, the policy will converge to the optimal policy regardless of whether the optimal policy is stochastic or deterministic.

- [x] When using softmax policy over action-preferences, assuming a tabular representation, the policy will converge to the optimal policy regardless of whether the optimal policy is stochastic or deterministic.

Correct

Correct. Action-preferences does not approach specific values like action-values do. They can be driven to produce a stochastic policy or deterministic policy.

6. What is the following objective, and in which task formulation? **1 / 1 point**

$$r(\pi) = \Sigma_s \mu(s) \Sigma_a \pi(a|s,\theta) \Sigma_{s',r} p(s',r|s,a) r$$

- ○ Average reward objective, episodic task
- ○ Discounted return objective, continuing task
- ○ Undiscounted return objective, episodic task
- ◉ Average reward objective, continuing task

Correct

Correct.

7. The following equation is the outcome of the policy gradient theorem. Which of the following is true about the policy gradient theorem? (Select all that apply) **1 / 1 point**

$$\nabla r(\pi) = \Sigma_s \mu(s) \Sigma_a \nabla \pi(a|s,\theta) q_\pi(s,a)$$

- [x] This expression can be converted into the following expectation over π:

  $$E_\pi[\nabla \ln \pi(A|S,\theta) q_\pi(S,A)]$$

  Correct
  Correct. In fact, this expression is normally used to perform stochastic gradient updates.

- [x] This expression can be converted into:

  $$E_\pi[\Sigma_a \nabla \pi(a|S,\theta) q_\pi(S,a)]$$

  In discrete action space, by approximating q_pi we could also use this gradient to update the policy.

  Correct
  Correct. The expression contains sum over actions, which can be computed for discrete actions. In the textbook, this is also called the all-actions method.

- [x] We do not need to compute the gradient of the state distribution μ.

  Correct
  Correct.

- [x] The true action value q_π can be approximated in many ways, for example using TD algorithms.

  Correct
  Correct.

8. Which of the following statements is true? (**Select all that apply**) **1 / 1 point**

- [x] Subtracting a baseline in the policy gradient update tends to reduce the variance of the update, which results in faster learning.

Correct
Correct.

- [x] To update the actor in Actor-Critic, we can use TD error in place of q_π in the Policy Gradient Theorem.

Correct
Correct. This is equivalent to using one-step state value and subtracting a current state value baseline.

- [x] The Actor-Critic algorithm consists of two parts: a parameterized policy — the actor — and a value function — the critic.

Correct
Correct.

- [ ] TD methods do not have a role when estimating the policy directly.

9. We usually want the critic to update at a faster rate than the actor. **1 / 1 point**

- (●) True
- ( ) False

Correct
Correct!

10. Consider the following state features and parameters θ for three different actions (red, green, and blue): **1 / 1 point**

$$\mathbf{X}(s) = \begin{bmatrix} 0.1 \\ 0.3 \\ 0.6 \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} 45 \\ 73 \\ 21 \\ 120 \\ 120 \\ \text{-}10 \\ \text{-}100 \\ 200 \\ \text{-}25 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix}0\\0\\0\end{matrix}}\right\} a_0 \\ \left.\vphantom{\begin{matrix}0\\0\\0\end{matrix}}\right\} a_1 \\ \left.\vphantom{\begin{matrix}0\\0\\0\end{matrix}}\right\} a_2 \end{array}$$

Compute the action preferences for each of the three different actions using linear function approximation and stacked features for the action preferences.

What is the action preference of a_0 (red)?

- ○ 33
- ○ 35
- ◉ 39
- ○ 37

Correct
Correct.

11. Which of the following statements are true about the Actor-Critic algorithm with softmax policies? (**Choose all that apply**)

1 / 1 point

- [x] The learning rate parameter of the actor and the critic can be different.

Correct

Correct! In practice, it is preferable to have a slower learning rate for the actor so that the critic can accurately critique the policy.

- [ ] The actor and the critic share the same set of parameters.
- [x] Since the policy is written as a function of the current state, it is like having a different softmax distribution for each state.

Correct

Correct!

- [ ] The preferences must be approximated using linear function approximation.

12. A Gaussian policy becomes deterministic in the limit $\sigma \longrightarrow 0$.

- (•) True
- ( ) False

Correct

Correct: As σ approaches 0, the values of the Gaussian policy approach the mean of the policy in a given state.