Policy Gradient Methods
Graded Quiz • 45 min

Due Feb 19, 12:59 AM MST

Policy Gradient Methods

TOTAL POINTS 12

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

□ **Policy gradient methods do gradient ascent on the policy objective.**

□ **Policy gradient methods use generalized policy iteration to learn policies directly.**

□ **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.**

□ **The policy gradient theorem provides a form for the policy gradient that does not contain the gradient of the state distribution \mu, which is hard to estimate.**

2. **Which of the following statements about parameterized policies are true? (Select all that apply)** 1 point

□ **The probability of selecting any action must be greater than or equal to zero.**

□ **For each state, the sum of all the action probabilities must equal to one.**

□ **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 point

○ **0.879**

○ **0.002**

○ **0.42**

○ **0.119**

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

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

□ **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.**

□ **Similar to epsilon-greedy policy, softmax policy cannot approach a deterministic policy.**

□ **It is used to represent a policy in discrete action spaces.**

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

1 point

□ **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.**

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

□ **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.**

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

1 point

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

○ **Discounted return objective, continuing task**

○ **Average reward objective, continuing task**

○ **Average reward objective, episodic task**

○ **Undiscounted return objective, episodic task**

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 point

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

□ **The true action value q_π can be approximated in many ways, for example using TD algorithms.**

□ **We do not need to compute the gradient of the state distribution μ.**

□ **This expression can be converted into the following expectation over π:**

$$\mathbb{E}_\pi[\nabla ln \pi(A|S,\theta) q_\pi(S,A)]$$

□ **This expression can be converted into:**

$$\mathbb{E}_\pi[\Sigma_a \nabla \pi(a|S,\theta) q_\pi(S,a)]$$

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

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

1 point

□ **TD methods do not have a role when estimating the policy directly.**

□ **Subtracting a baseline in the policy gradient update tends to reduce the variance of the update, which results in faster learning.**

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

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

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

○ True

○ False

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

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

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_1 (green)?

○ 42

○ 40

○ 35

○ 32

11. Which of the following statements are true about the Actor-Critic algorithm with softmax policies? (Choose all that apply) 1 point

□ Since the policy is written as a function of the current state, it is like having a different softmax distribution for each state.

□ The preferences must be approximated using linear function approximation.

□ The learning rate parameter of the actor and the critic can be different.

□ The actor and the critic share the same set of parameters.

12. **Which of the following is an advantage of Gaussian policy parameterization over discretizing the action space? (Select all that apply)** 1 point

□ There might not be a straightforward way to choose a discrete set of actions.

□ Even if the true action set is discrete, but very large, it might be better to treat them as a continuous range.

□ Continuous actions also allow learning to generalize over actions.

□ Gaussian policies are differentiable, whereas policies over discretized actions are not.

□ I, **Dhawal Gupta**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code

Save Submit