## Congratulations! You passed!

0.119

**⊘** Correct

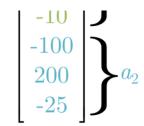
Grade received 100% To pass 80% or higher

Go to next item

**Policy Gradient Methods** Total points 12 1. Which of the following is true about policy-based methods? (Select all that apply) 1 / 1 point Policy-based methods can be applied to continuous action space domains. **⊘** Correct Correct. By parameterizing a policy to represent a probability distribution such as Gaussian, it can be applied to continuous action space domains. Policy-based methods can learn an optimal policy that is stochastic. **⊘** Correct Correct. It can learn a stochastic optimal policy, such as the soft-max in action preferences. Policy-based methods are useful in problems where the policy is easier to approximate than actionvalue functions. Correct. For example in the Mountain Car problem a good policy is easy to represent whereas the value function is complex. Policy-based methods allow smooth improvement in the policy without drastic changes. **⊘** Correct Correct. As the policy parameters change the action probabilities change smoothly, but with valuebased methods a small change in action-value function can drastically change the action probabilities. 2. Which of the following statements about parameterized policies are true? (Select all that apply) 1/1 point ☐ The policy must be approximated using linear function approximation.  $\hfill \square$  The function used for representing the policy must be a softmax function. The probability of selecting any action must be greater than or equal to zero. Correct! This is one of the conditions for a valid probability distribution. For each state, the sum of all the action probabilities must equal to one. Correct! This condition is necessary for the function to be a valid probability distribution. 3. Assume you're given the following preferences  $\,h_1=44$ ,  $\,h_2=42$ , and  $\,h_3=38$ , corresponding to three 1/1 point 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? 0.42 0.879 0.002

✓ CorrectCorrect.

$oldsymbol{ abla}$ The true action value $q_\pi$ can be approximated in many ways, for example using TD algorithms.	
○ Correct     Correct.	
This expression can be converted into the following expectation over $\pi$ : $\mathbb{E}_{\pi}[\nabla ln\pi(A S,\theta)q_{\pi}(S,A)]$	
<ul> <li>Correct</li> <li>Correct. In fact, this expression is normally used to perform stochastic gradient updates.</li> </ul>	
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.	
<ul> <li>Correct</li> <li>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.</li> </ul>	
<ul> <li>8. Which of the following statements is true? (Select all that apply)</li> <li>The Actor-Critic algorithm consists of two parts: a parameterized policy — the actor — and a value</li> </ul>	1/1 point
function — the critic.  Orrect Correct.	
$\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $	
<ul> <li>Correct</li> <li>Correct. This is equivalent to using one-step state value and subtracting a current state value baseline.</li> </ul>	
<ul> <li>Subtracting a baseline in the policy gradient update tends to reduce the variance of the update, which results in faster learning.</li> <li>Correct</li> </ul>	
Correct.	
<ul> <li>To train the critic, we must use the average reward version of semi-gradient TD(0).</li> <li>True</li> </ul>	1/1 point
<ul><li>● False</li><li>✓ Correct</li></ul>	
Correct. We can use any state-value learning algorithm.	
10. Consider the following state features and parameters $\theta$ for three different actions (red, green, and blue):	1 / 1 point
$\begin{bmatrix} 45 \\ 73 \\ 21 \\ 120 \end{bmatrix} a_0$	
$\mathbf{X}(s) = \begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} 120 \\ 120 \\ 100 \end{bmatrix} \boldsymbol{a}_1$	



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

	and stacked reatures for the action preferences.
	What is the action preference of $a_0$ (red)?
	③ 39
	○ 33
	○ 37
	○ 35
11.	, Which of the following statements are true about the Actor-Critic algorithm with softmax policies? (Choose all that apply)
	✓ 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 preferences must be approximated using linear function approximation.
	Since the policy is written as a function of the current state, it is like having a different softmax distribution for each state.
	The actor and the critic share the same set of parameters.
12	, A Gaussian policy becomes deterministic in the limit $\sigma  o 0.$
	(a) True
	○ False
	$\bigcirc$ Correct Correct: As $\sigma$ approaches 0, the values of the Gaussian policy approach the mean of the policy in a given state.