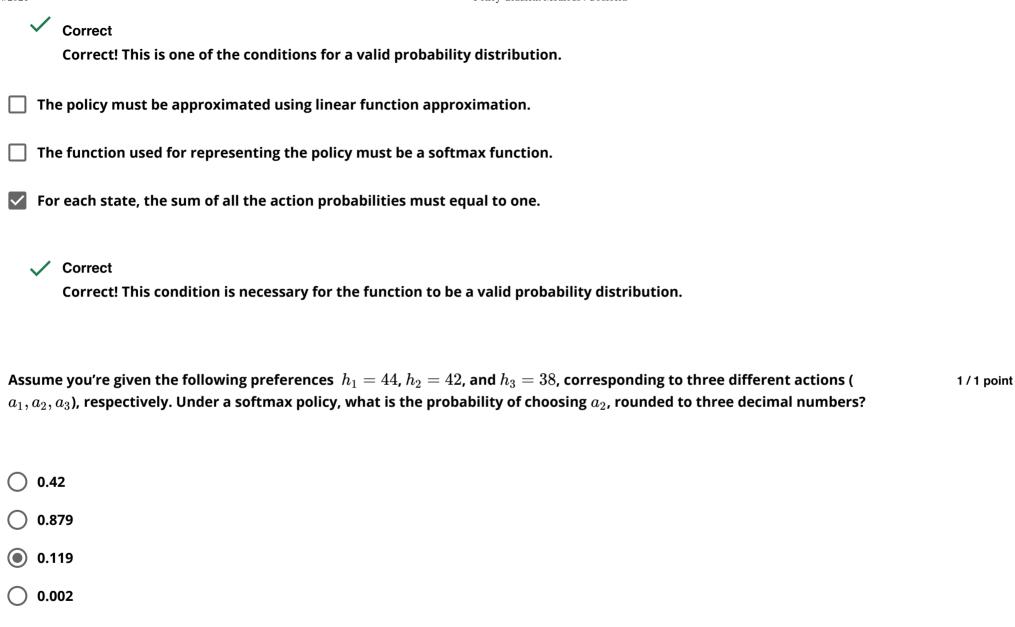
2.

Which of the following is true about policy gradient methods? (Select all that apply)	1/1 point
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.	
\checkmark Correct Correct. We derived this stochastic update by multiplying and dividing by $\pi(A S)$.	
Policy gradient methods use generalized policy iteration to learn policies directly.	
Policy gradient methods do gradient ascent on the policy objective.	
 Correct Correct. Policy gradient methods maximize the policy objective, and hence perform gradient ascent. 	
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.	
✓ Correct Correct.	
Which of the following statements about parameterized policies are true? (Select all that apply)	1 / 1 point
The probability of selecting any action must be greater than or equal to zero.	



Correct!

4.	4. Which of the following is true abo	out softmax policy? (Select all that apply)	1/1 point
	It can be parameterized by a form a softmax policy.	ny function approximator as long as it can output scalar values for each available actio	n, to
	CorrectCorrect. It can use any full	nction approximation from deep artificial neural networks to simple linear features.	
	Similar to epsilon-greedy poli	licy, softmax policy cannot approach a deterministic policy.	
	It is used to represent a polic	cy in discrete action spaces.	
	✓ Correct!		
	It cannot represent an optim dominates others.	nal policy that is stochastic, because it reaches a deterministic policy as one action prefe	erence
5.	 What are the differences between (Select all that apply) 	en using softmax policy over action-values and using softmax policy over action-prefere	n ic@\$3 333333333333 / 1 point
		over action-values, assuming a tabular representation, the policy will converge to the oper the operation that the operation is stochastic or deterministic.	otimal
	This should not be selected	ed	

Incorrect. The action-value estimates would converge to the true values which would differ by a finite amount, and each action would have fixed probabilities other than 0 or 1. Softmax policy over action-values is unlikely to be the optimal policy and may never be deterministic.

- 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.

- When using softmax policy over action-values, even if the optimal policy is deterministic, the policy may never approach a deterministic policy.
- 6. What is the following objective, and in which task formulation?

1/1 point

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

- O Discounted return objective, continuing task
- Oundiscounted return objective, episodic task
- Average reward objective, continuing task
- Average reward objective, episodic 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)

$$\nabla r(\pi) = \sum_{s} \mu(s) \sum_{a} \nabla \pi(a|s,\theta) q_{\pi}(s,a)$$

This expression can be converted into:

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

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.

- The true action value q_{π} can be approximated in many ways, for example using TD algorithms.
 - Correct
- We do not need to compute the gradient of the state distribution μ .

Correct

Correct.

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

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

✓ Correct

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

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

1/1 point

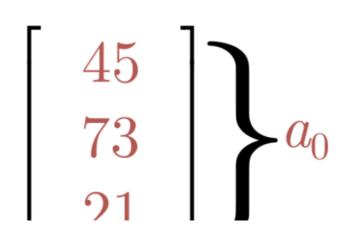
- **TD** methods do not have a role when estimating the policy directly.
- lacktriangle To update the actor in Actor-Critic, we can use TD error in place of q_π in the Policy Gradient Theorem.
 - ✓ Correct
 Correct. This is equivalent to using one-step state value and subtracting a current state value baseline.
- The Actor-Critic algorithm consists of two parts: a parameterized policy the actor and a value function the critic.
 - Correct

- Subtracting a baseline in the policy gradient update tends to reduce the variance of the update, which results in faster learning.
- Correct
- 9. We usually want the critic to update at a faster rate than the actor.

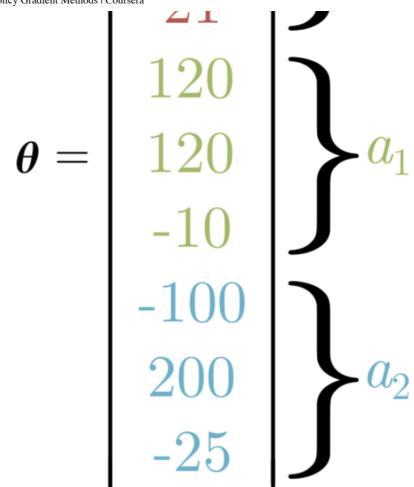
1/1 point

- True
- False
 - ✓ Correct!
- 10. Consider the following state features and parameters θ for three different actions (red, green, and blue):

1/1 point



$$\mathbf{X}(s) = \begin{bmatrix} 0.1 \\ 0.3 \\ 0.6 \end{bmatrix}$$



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_2 (blue)?

- 35
- 39
- **42**

\bigcirc	37
\bigcup	51

/	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

- The preferences must be approximated using linear function approximation.
- The actor and the critic share the same set of parameters.
- 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.
- Since the policy is written as a function of the current state, it is like having a different softmax distribution for each state.
 - ✓ Correct!
- 12. Which of the following is an advantage of Gaussian policy parameterization over discretizing the action space? (Select all that apply)

0/1 point

14/2021	Toney Gradient Methods (Coursela
Even if the t	rue action set is discrete, but very large, it might be better to treat them as a continuous range.
✓ Correct Correct.	
▼ There might	not be a straightforward way to choose a discrete set of actions.
	Selecting a discrete set of actions that results in good performance is problem dependent. Maybe we need s of actions. Maybe it is state dependent!
Continuous	actions also allow learning to generalize over actions.
✓ Correct Correct.	
Gaussian po	licies are differentiable, whereas policies over discretized actions are not.

Incorrect. Consider the softmax policy, which is a policy over discrete actions and differentiable.

X This should not be selected