Congratulations! You passed!

TO PASS 80% or higher

Keep Learning

GRADE
100%

Policy Gradient Methods

TOTAL POINTS 18

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

- [x] **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.

- [x] **Policy-based methods are useful in problems where the policy is easier to approximate than action-value functions.**

  Correct
  Correct. For example in the Mountain Car problem a good policy is easy to represent whereas the value function is complex.

- [x] **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 value-based methods a small change in action-value function can drastically change the action probabilities.

- [x] **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.

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 function used for representing the policy must be a softmax function.
- [ ] The policy must be approximated using linear function approximation.

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

- ( ) 0.879
- (•) 0.119
- ( ) 0.002
- ( ) 0.42

Correct

Correct!

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

- [ ] Similar to epsilon-greedy policy, softmax policy cannot approach a deterministic policy.
- [x] It is used to represent a policy in discrete action spaces.

Correct

Correct!

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

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-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.

- [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.

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

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

○ Discounted return objective, continuing task

◉ Average reward objective, continuing task

○ Average reward objective, episodic task

○ Undiscounted return objective, episodic task

✓ Correct
Correct.

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

☑ 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.

☑ 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 $\pi(A|S)$.

☐ 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.

8. 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:

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

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

- [x] This expression can be converted into:

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

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

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

  ✓ Correct
  Correct.

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

  ✓ Correct
  Correct.

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

  ✓ 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.

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

10. **To train the critic, we must use the average reward version of semi-gradient TD(0).** 1 / 1 point

- ○ True
- ◉ False

✓ Correct
Correct. We can use any state-value learning algorithm.

11. **Question 11 ~ 13: Consider the following state features and parameters θ for three different actions (red, green, and blue):** 1 / 1 point

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

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_0 (red)?

- ○ 35
- ○ 33
- ○ 37
- ◉ 39

✓ Correct
Correct.

12. **What is the action preference of a_1 (green)?** 1 / 1 point

- 32
- 42
- 35
- 40

Correct
Correct.

13. What is the action preference of a_2 (blue)? 1 / 1 point

- 42
- 35
- 39
- 37

Correct
Correct.

14. 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.

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

- True
- False

Correct
Correct.

16. Which one is a reasonable parameterization for a Gaussian policy? 1 / 1 point

- ○ μ: a linear function of parameters, σ: a linear function of parameters
- ○ μ: the exponential of a linear function of parameters, σ: a linear function of parameters.
- ◉ μ: a linear function of parameters, σ: the exponential of a linear function of parameters.

✓ Correct
Correct!

17. A Gaussian policy becomes deterministic in the limit $\sigma \to 0$. 1 / 1 point

- ◉ True
- ○ False

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

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

- ☑ Continuous actions also allow learning to generalize over actions.

✓ Correct
Correct!

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

✓ Correct
Correct! Selecting a discrete set of actions that results in good performance is problem dependent. Maybe we need hundreds of actions. Maybe it is state dependent!

- ☐ Gaussian policies are differentiable, whereas policies over discretized actions are not.
- ☑ Even if the true action set is discrete, but very large, it might be better to treat them as a continuous range.

✓ Correct
Correct!