# Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 8

Solution by : Amirhossein Asadi

## Q1

**In PPO's clipped objective, if the probability ratio:**

$$\frac{\pi_{\theta'}(a|s)}{\pi_{\theta}(a|s)} = 1.5$$

**and the clipping threshold:**

$$\epsilon = 0.2$$

**what is the value of the clipped ratio?**

A) 1.2
B) 1.3
C) 1.5
D) 1.7

**Correct Answer: A**

The clipped ratio is calculated as:

$$\text{Clipped ratio} = \text{clip}\left(\frac{\pi_{\theta'}(a|s)}{\pi_{\theta}(a|s)}, 1-\epsilon, 1+\epsilon\right)$$

$$\text{Clipped ratio} = \text{clip}(1.5, 0.8, 1.2)$$

Since $1.5$ is greater than $1.2$, the clipped ratio is:

$$\boxed{1.2}$$

## Q2

**Increasing the entropy weight $\alpha$ in SAC's objective would most likely:**

A) Reduce exploration and prioritize immediate rewards.
B) Increase exploration and penalize low-entropy policies.
C) Speed up Q-function convergence.
D) Decrease robustness to environment noise.

**Correct Answer: B**

Increasing the entropy weight $\alpha$ in SAC's objective increases exploration by encouraging the policy to have higher entropy, which avoids deterministic policies and favors exploration. This penalizes low-entropy policies and ensures the agent explores more diverse actions.

## Q3

**In the SAC algorithm, which step ensures the policy stays close to the exponential of the Q-values?**

A) Q-function update using: $r(s, a) + \gamma * E[V(s')]$
B) Value function regression on mean-squared error.
C) Policy update via KL divergence minimization.
D) Exponential moving average of $\bar{\psi}$.

**Correct Answer: D**

The exponential moving average of $\bar{\psi}$ ensures that the policy stays close to the exponential of the Q-values by stabilizing the target value function and reducing variance during training.

## Q4

**PPO uses importance sampling corrections because it is an off-policy algorithm.**

A) True
B) False

**Correct Answer: B**

The correct answer is **False**. PPO is an on-policy algorithm, meaning it updates the policy using data from the current policy. Therefore, it does not require importance sampling corrections, which are typically used in off-policy algorithms like Q-learning or DQN.

## Q5

**The SAC soft Bellman equation includes a term:**

$$-\log \pi(a'|s')$$

**What does this term represent?**

A) Discount factor
B) Entropy regularization
C) Advantage function
D) Value function error

**Correct Answer: B**

In SAC, the term $-\log \pi(a'|s')$ represents entropy regularization. The entropy term encourages the agent to explore a variety of actions instead of converging too quickly to a deterministic policy. This helps prevent the policy from becoming too deterministic, which could lead to suboptimal behavior in uncertain environments. The entropy regularization term in SAC thus strikes a balance between exploration and exploitation, promoting better exploration to find more optimal solutions.

## Q6

**During PPO training, you observe large policy updates destabilizing learning. Which adjustment would best mitigate this?**

A) Decrease the KL divergence constraint $\epsilon$.
B) Remove the clip function entirely.
C) Switch to an off-policy algorithm.
D) Increase $\epsilon$ in the clip function.

**Correct Answer: A**

If PPO's policy updates are too large, it may be because the KL divergence constraint is too large. Reducing the value of $\epsilon$ (the constraint on the size of policy updates) will allow for smaller, more stable updates and prevent destabilizing the learning process. This helps to keep the policy closer to the old policy, preventing drastic changes.

## Q7

**In Soft Actor-Critic (SAC), why is the entropy term:**

$$\alpha \mathbb{H}(\pi(\cdot|s_t))$$

**included in the objective function $J(\pi)$?**

**Correct Answer: To encourage exploration and prevent the policy from becoming deterministic too quickly.**

In SAC, the entropy term is included in the objective function to strike a balance between exploration and exploitation. The entropy of the policy $\mathbb{H}(\pi(\cdot|s_t))$ measures the uncertainty or randomness of the actions taken by the agent. When the entropy term is added to the objective, it encourages the agent to explore a wider range of actions, rather than converging to a deterministic policy too quickly. This is particularly useful in environments where the optimal policy is not immediately obvious, and exploration is necessary to discover good strategies. The weight $\alpha$ in front of the entropy term controls the trade-off between maximizing expected reward and maintaining high entropy (which favors exploration). A higher $\alpha$ puts more emphasis on exploration, whereas a lower $\alpha$ prioritizes exploiting known strategies. By including this entropy term, SAC ensures that the policy remains flexible, allowing the agent to explore more diverse strategies and avoid getting stuck in local optima.