What is the purpose of using dual Q-networks in SAC and SAC++?
A) To minimize the computational overhead during each update step.
B) To reduce the positive bias that can arise in policy updates, improving training stability.
C) To enable both on-policy and off-policy updates simultaneously.
D) To ensure that the policy network learns directly from the critic network's feedback without delay.

Correct Answer: B

Which of the following best describes how the maximum entropy framework improves exploration in SAC?
A) By forcing the policy to always select the action with the highest expected reward.
B) By regularizing the value function, ensuring that it remains close to the target value.
C) By encouraging the policy to maintain stochasticity, leading to diverse behavior and better exploration of the state space.
D) By penalizing actions that deviate from an expert policy, ensuring more stable learning.

Correct Answer: C

What is the role of the temperature parameter α in the maximum entropy objective of SAC?
A) It directly controls the trade-off between maximizing reward and entropy, impacting exploration and exploitation.
B) It ensures that the policy becomes deterministic over time by reducing stochasticity gradually.
C) It acts as a regularization term that prevents overfitting to specific policies during training.
D) It adjusts the learning rate dynamically based on the variance of the Q-values in each training step.

Correct Answer: A

In SAC++, what effect does the automatic tuning of the temperature parameter α have on the policy's behavior when the observed entropy is higher than the target entropy?
A) It increases the temperature to promote even more exploration.
B) It decreases the temperature to reduce the randomness of actions, promoting more exploitation.
C) It keeps the temperature constant to maintain stability in the learning process.
D) It resets the temperature to a predefined value based on the task's complexity.

Correct Answer: B