Reinforcement Learning — Boxed Questions (Large Font)

Q1. (Multi-select) Which algorithmsong to the policy-gradient
family?
(A) Proximal Policy Optimization (PPO)
(B) Q-learning
(C) REINFORCE
(D) Deep Q-Network (DQN)
Q2. An environment is considered partially observable when:
(A) The agent can observe the full state
(B) Observations lack some hidden variables
(C) The reward signal is stochastic
(D) Transition dynamics are deterministic
Q3. In Q-learning, the key difference from SARSA is:
(A) On-policy vs. off-policy nature
(B) Use of neural networks
(C) Continuous action space support
(D) Model-based planning

Q4. The Advantage term in Actor–Critic methods is calculated as:
(A) State-value minus action-value
(B) Action-value minus state-value
(C) Reward minus entropy
(D) Policy-gradient estimate
Q5. (Multi-select) Techniques to stabilize deep-RL training include:
(A) Experience replay
(B) Target networks
(C) Layer normalization
(D) Early stopping
Q6. A discount factor γclose to 0 emphasizes:
(A) Immediate rewards
(B) Long-term rewards
(C) Exploration
(D) Deterministic policies
Q7. The exploration–exploitation dilemma refers to:

- (A) Balancing policy and value networks
- (B) Choosing between trying new actions and using known good actions
- (C) Data augmentation
- (D) Hyper-parameter tuning