Online and Reinforcement Learning (2025)
Home Assignment 7

Davide Marchi 777881

Contents

1 Short Questions

1. **True.** *Justification:* In a finite average-reward MDP with a finite diameter the MDP is communicating (in fact, ergodic if every state is reachable from every other under some policy). This implies that the long-run average reward (gain) is independent of the starting state.

2. **False.** *Justification:* Finite diameter guarantees that there exists a policy which can reach any state from any other in a finite expected number of steps. However, this does not mean that every arbitrary choice of actions will eventually reach every state.

3. **False.** *Justification:* In an ergodic MDP the optimal gain is unique (i.e., independent of the initial state), but the optimal bias function is determined only up to an additive constant. Hence, the bias is not uniquely defined in an absolute sense.

4. **False.** *Justification:* A PAC-MDP algorithm guarantees that the number of ε-bad (i.e., non-ε-optimal) steps is bounded with high probability. However, it does not imply that there exists a finite time after which every subsequent policy is ε-optimal; occasional exploration may still yield suboptimal actions.

2 Offline Evaluation of Bandit Algorithms - the Practical Part

3 Grid-World: Continual and undiscounted

4 An Empirical Evaluation of UCB Q-learning