Reinforcement Learning
Exercise 4

Mathias Niepert, Vinh Tong

May 3, 2024

1 Monte Carlo Methods vs Dynamic Programming (3P)

a) What are advantages of Monte Carlo methods over dynamic programming? Mention at least two. (2P)

b) Give an example environment where you would use a Monte Carlo method to learn the value function rather than using dynamic programming. Explain why. (1P)

2 Monte Carlo ES for blackjack (6P)

In this exercise we use the blackjack environment from gym (https://www.gymlibrary.dev/environments/toy_text/blackjack/). The code template can be found on Ilias in *ex04-mc/ex04-mc.py*. It has been tested with gym version 0.18.0 (but should also be stable with version 0.18.3).

a) Consider the version of blackjack introduced in the lecture (Example 5.1 from Sutton and Barto). Implement first-visit Monte Carlo prediction (lecture 4 slide 13) for the given policy: stick if sum ≥ 20, else hit. Try to reproduce the figures on slide 16. (3P)

b) Implement Monte Carlo ES (slide 23) and obtain the optimal policy and state-value function for blackjack. Output the policy every 100,000 iterations (e.g. as 2 tables, one with usable ace and one without usable ace). We recommend an optimistic initialisation of Q to improve results. Let it run for at least 500,000 iterations. Letting it run until convergence might take a very long time, so intermediate results are okay! (3P)