Exercise 6

Kai Schneider

June 8, 2021

Task 1 Planning and Learning

a.)

- In the first phase Dyna-Q+ performs better due to the additional reward given by exploring the environment
- After changing the environment, Dyna-Q+ still performs better because it keeps on exploring and thus is able to find the new shortcut. In contrast, Dyna-Q is unable to find the shorter path because the already trained Q-function always chooses the "left" path. This way the reward rate stays the same, while the one of Dyna-Q+ increases after a certain time.

Dyna-Q+ improves exploring the environment as well as adapting to changing circumstances.

b.)

In an stochastic environment the agent can't determine transition and reward completely because it isn't deterministic
One solution to this problem would be to keep track of the rewards given by die (s,a)-pairs in the environment in form of probabilities ($\sim P(r|s,a,s')$). We than can use this information during the planning phase.
If our environment is not only non-deterministic, but also changing, this method could be problematic. Our "old" probabilities would no longer be useful for approximating our real environment. A solution could be some sort of decay to "forget" too old data. By using only a few newer samples instead of the whole data for calculating our probabilities we can adapt to a changing environment.

Task 2

a.)

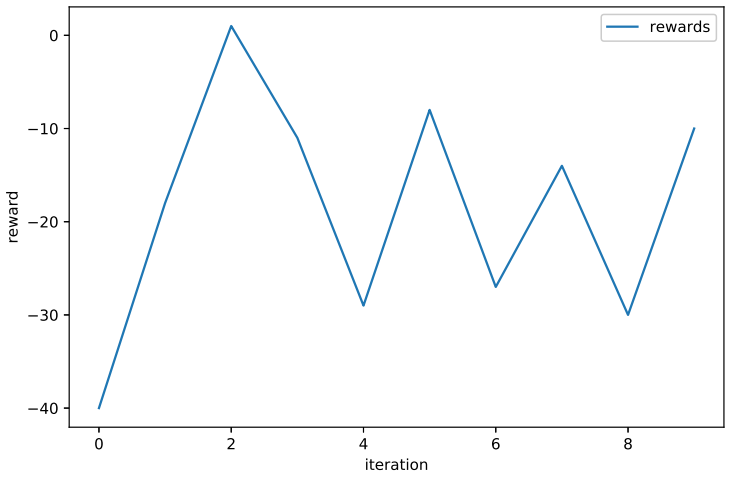

Figure 1: average reward over the iterations

b.)

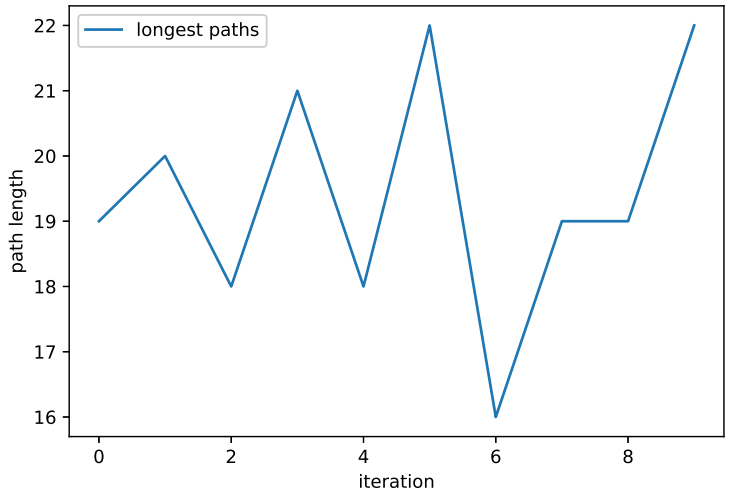

Figure 2: plot of length of the longest path over the iterations