Homework 4

Daniel Frank

May 19, 2017

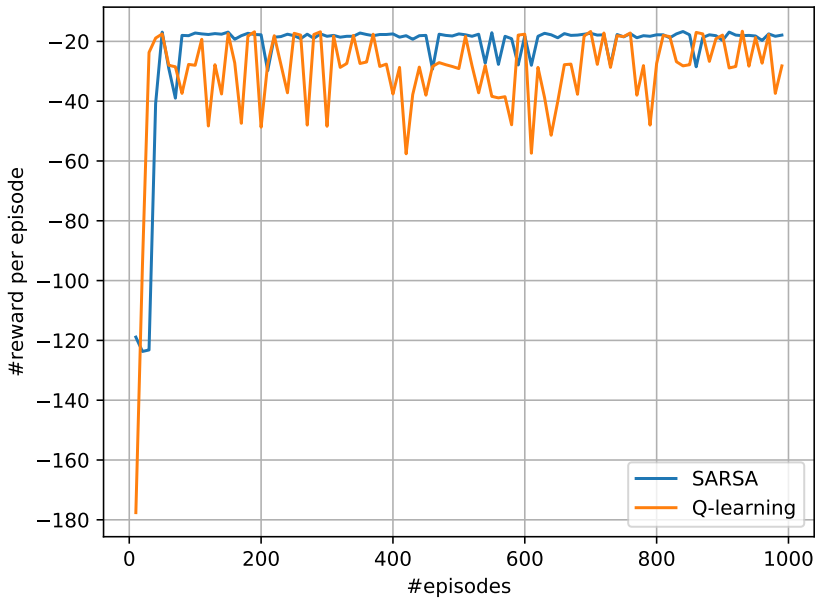

Figure 1: SARSA vs. Q-learning after 1000 episodes

1 Cliff Walking

2a)

For a fixed $\epsilon = 0.1$ the smoothed result is shown in figure 1.

The resulting target policy π is shown in figure 2. I do not know why the behaviour is not matching the one in the book, where SARSA represents the save way and Q-learning the risky way. Do you have any idea? Could the fixed ϵ be the problem or do I something wrong during the algorithm?

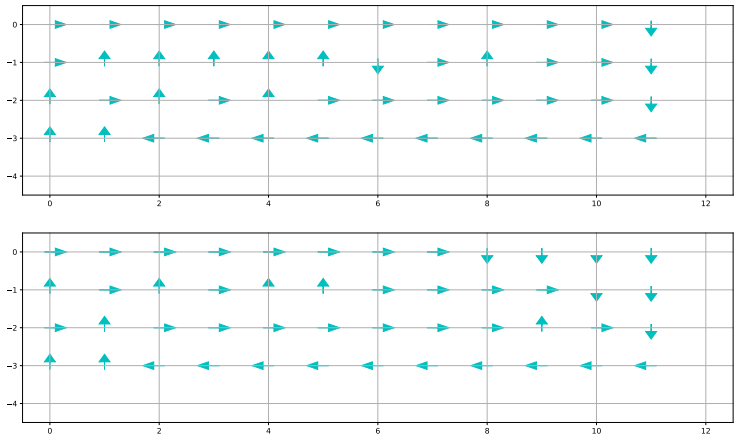

Figure 2: (top)target policy of SARSA (bottom)target policy Q-learning, both after 1000 episodes.

2b)

ϵ changes like

```
epsilonint = 1
epsilon.append(1/float(j+1) *epsilonint)
```

. Where j is the number of episodes. In figure 3 the result is shown.

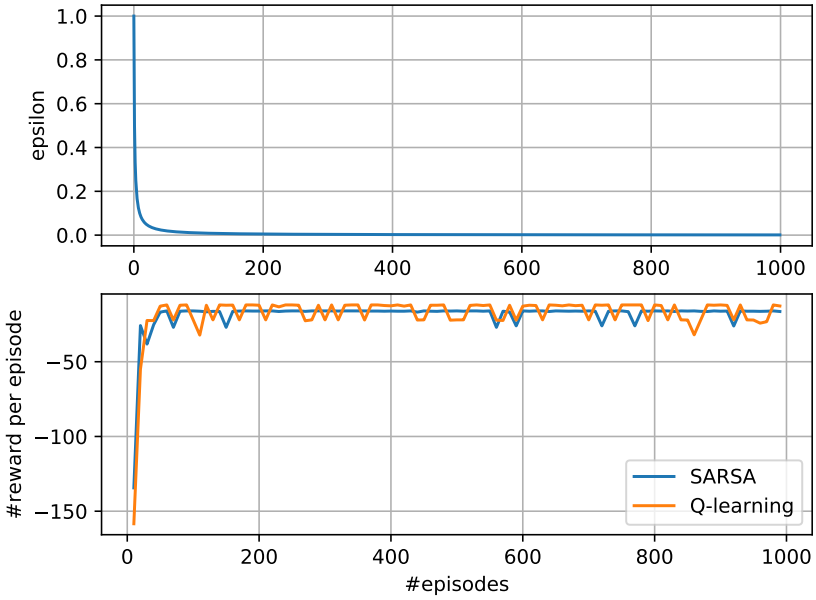

Figure 3: SARSA vs. Q-learning for dynamic ϵ

2c)

In this part $\lambda = [0.9, 0.8, 0.7, 0.5, 0.2]$ is not fixed anymore. The resulting plot is shown in figure 4

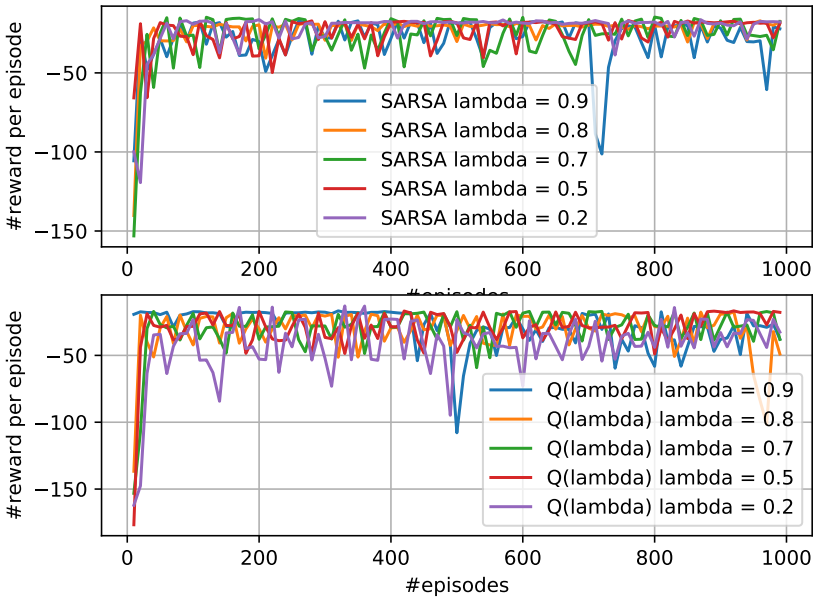

Figure 4: Different values for λ but fixed $\epsilon = 0.1$.