Temporal Difference methods

Exercise 1

$$\begin{aligned} V(B) &= V(B) + \alpha(R + \gamma V(A) - V(B)) \\ &= V(B) + \alpha(R + \gamma V(C) - V(B)) \\ &= 0.5 + 0.1(0.0 + 0.5 - 0.5) \\ &= 0.5 \end{aligned} \tag{1}$$

Analogously $V(A) = V(B) = V(C) = V(D) = V(E)$, if going left at state E and right at state A. Else:

$$\begin{aligned} V(E) &= V(E) + \alpha(R + \gamma V(terminal) - V(E)) \\ &= 0.5 + 0.1(1.0 + 0 - 0.5) \\ &= 0.55 \end{aligned} \tag{2}$$

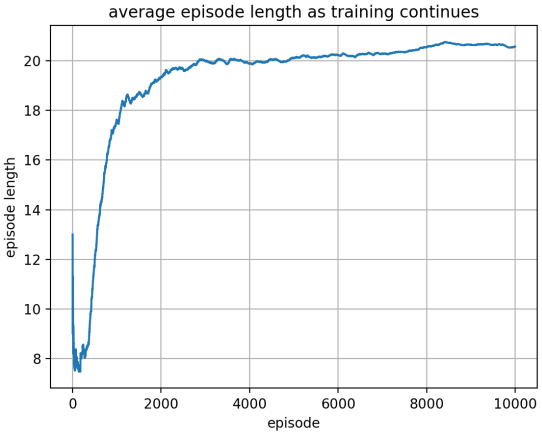
$$\begin{aligned} V(A) &= V(A) + \alpha(R + \gamma V(terminal) - V(A)) \\ &= 0.5 + 0.1(0 + 0 - 0.5) \\ &= 0.45 \text{ , what we observe} \end{aligned} \tag{3}$$

Thus, what we observe at the first episode is that it terminated after last visiting state A. (...and obviously not E, otherwise the value of A would be unchanged and the value of E become higher - one single episode cannot end in two terminal states). The exact difference of the value of A is 0.

Exercise 2

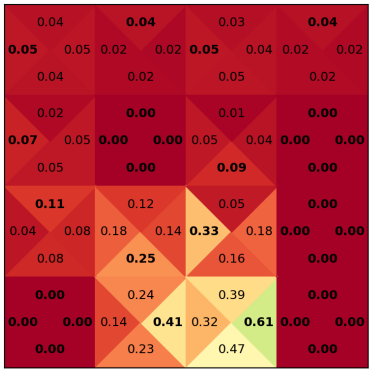
(a)

←	↑	←	↑
←	←	↓	←
↑	↓	←	←
←	→	→	←

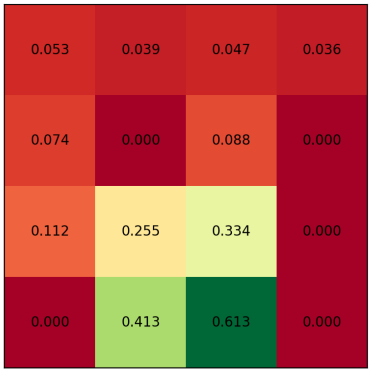


(a) SARSA

Figure 1: Episode lengths



(a) Action values

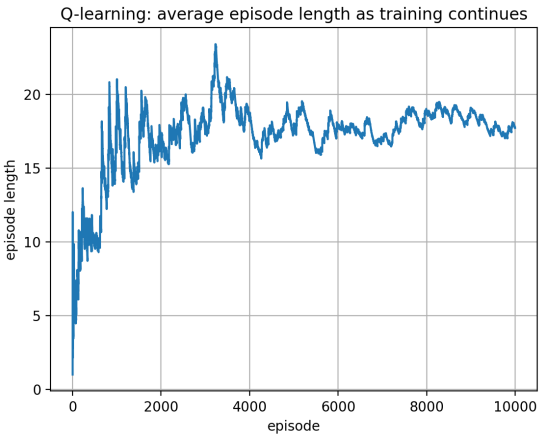


(b) State values

Figure 2: SARSA

(b)

→ ↑ ← ↑
← ← ← ←
↑ ↓ ← ←
← → ↓ ←



(a) Q-Learning

Figure 3: Episode lengths

(a) Action-values

(b) State values

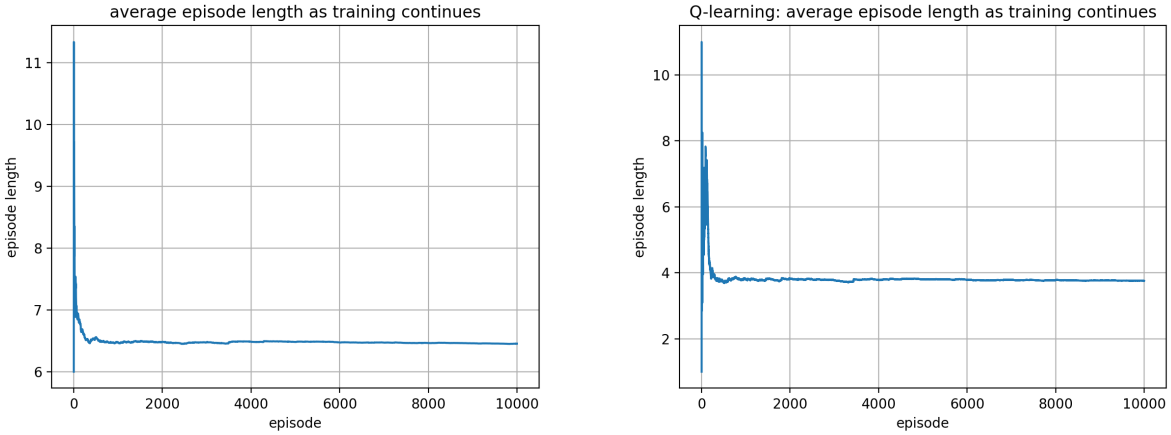
Figure 4: Q-Learning

Even though it also find the optimal Q-values (updating with the greedy action), the Q-learning algorithm doesn't generate it's steps following this policy (taking ϵ-greedy

actions). This behaviour can be seen in the episode length, being much more unstable than in the on-line case with SARSA. Apparently, similar to the cliff-example, the off-line Q-learning by taking ϵ-greedy actions following the optimal policy ends up at holes, while SARSA follows more save policy and avoids the holes better even when taking ϵ-greedy actions.

(c)

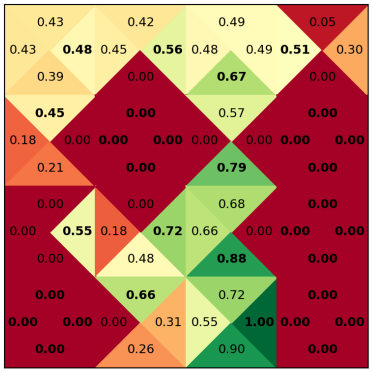
```
Running sarsa...
→ → ↓ ←
↑ ← ↓ ←
→ → ↓ ←
← ↑ → ←

Running qlearning
↓ ← ← ←
↓ ← ↑ ←
→ ↓ ↓ ←
← → → ←
```



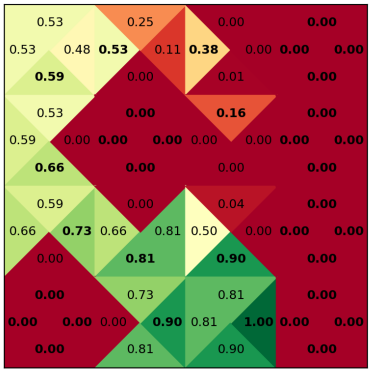
(a) SARSA episode length

(b) Q-Learning episode length

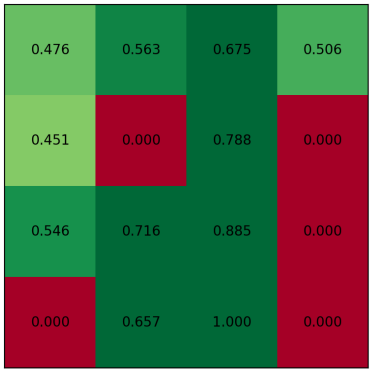
Figure 5: Episode lengths on non-slippery surface



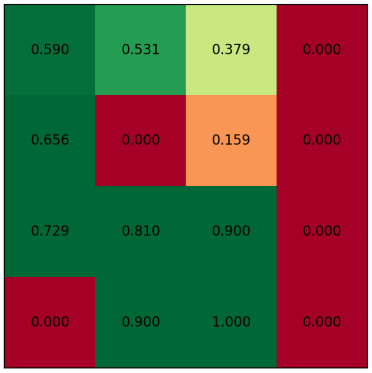
(a) SARSA: Action values



(b) Q-learning: Action values



(c) SARSA: State values

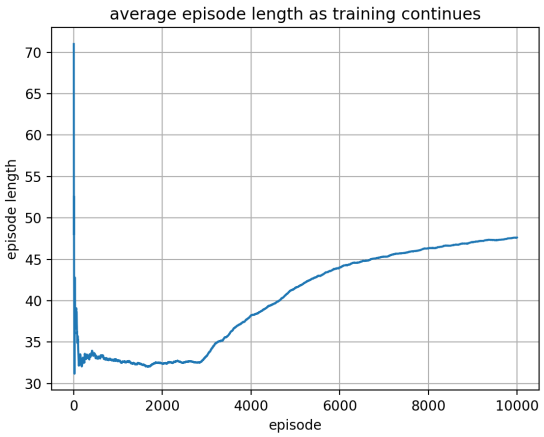


(d) Q-Learning: State values

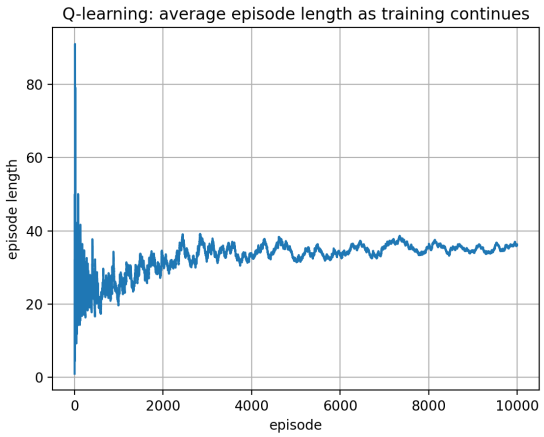
Figure 6: Action-values and state values on non-slippery surface

(d)

```
Running sarsa...
↑ → → → → → → →
↑ ↑ ↑ ↑ → → ↓ ↓
↑ ↑ ← ← → ↑ → ←
← ← ↑ ↓ ← ← → →
→ ↑ ← ← → ↓ ↑ →
← ← ← ↓ ↑ ← ← ↓
↑ ← ↓ ← ← ← ← →
↑ ↑ ← ← → ↓ ↓ ←
Running qlearning
↑ → → → → → → →
↑ → ↑ ↑ → → → ↓
↑ → ← ← → ↑ ↓ ↓
↑ ↑ ↑ → ← ← → ←
→ ← ↑ ← → ↓ ↑ →
→ ← ← ↓ ↓ ← ← →
← ← → → ← ← ← →
↑ ← ← ← ↓ ↑ ↓ ←
```

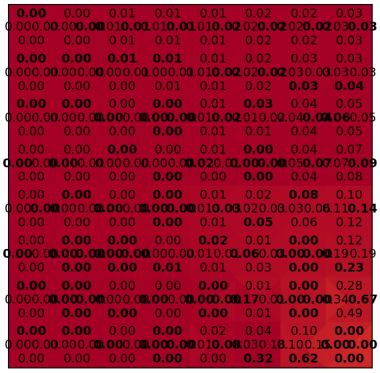


(a) SARSA episode length

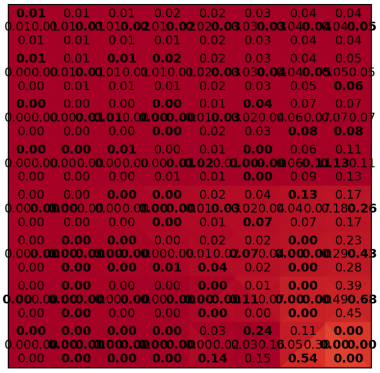


(b) Q-Learning episode length

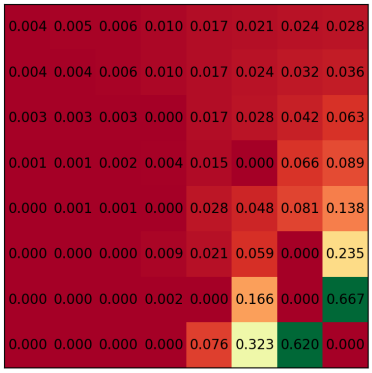
Figure 7: Episode lengths on 8x8 surface



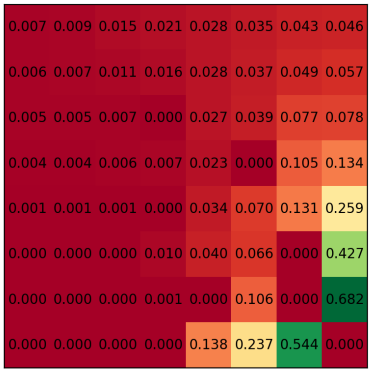
(a) SARSA: Action values



(b) Q-learning: Action values



(c) SARSA: State values



(d) Q-Learning: State values

Figure 8: Action-values and state values on 8x8 surface