Exercise 5

Task 1 - Random Walk

- The first episode terminates in the left state.
- The states are not yet visited or the ones visited are updated with TD error = 0 (→ the values are not changing at all.)
- $V_{t+1}(A) = V_t(A) + \alpha * (R_{t+1} + \gamma V(terminal) - V_t(A))$ in our case: $0.5 + 0.1 * (0 + 0 - 0.5) = 0.45$

Task 2 - Sarsa and Q-learning on the FrozenLake

a)

see Figure 1 & Figure 2 & Figure 3.

←	↑	↓	↑
←	←	←	←
↑	↓	←	←
←	→	↑	←

Table 1: Sarsa policy

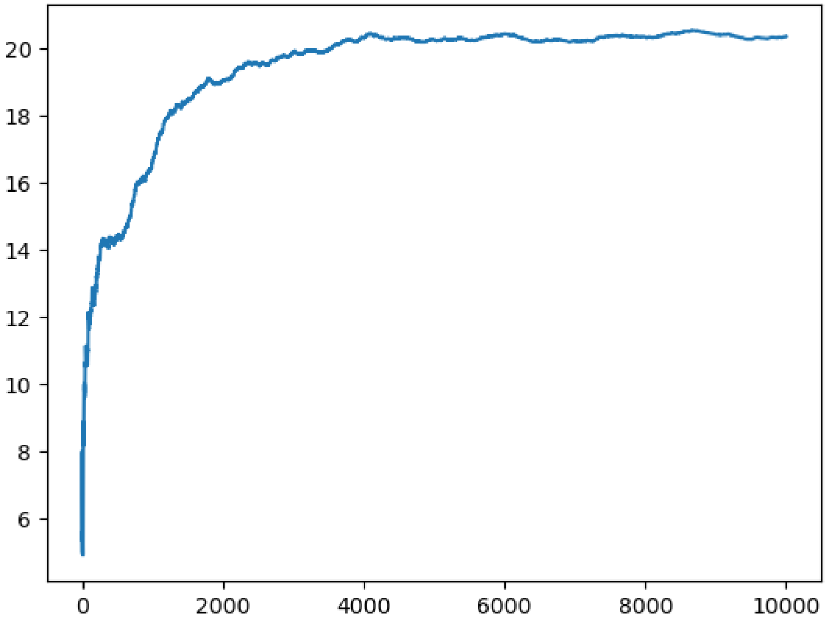


Figure 1: Sarsa training length

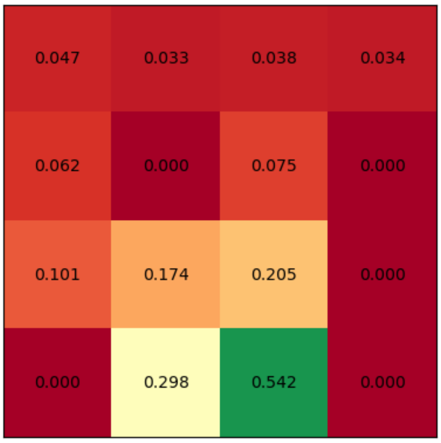


Figure 2: Sarsa V

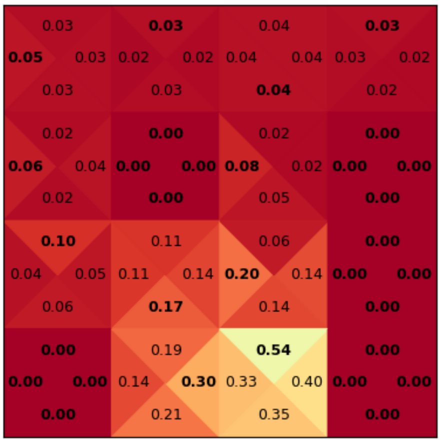


Figure 3: Sarsa Q

b)

Sarsa follows a more saver policy (more exploration) then q-learning (more exploitation). Same as in the cliff example in the lecture Q-learning takes the optimal path which could end up in holes.

see Figure 4 & Figure 5 & Figure 6.

↓	↑	←	↑
←	←	←	←
↑	↓	←	←
←	→	→	←

Table 2: Q-Learning policy

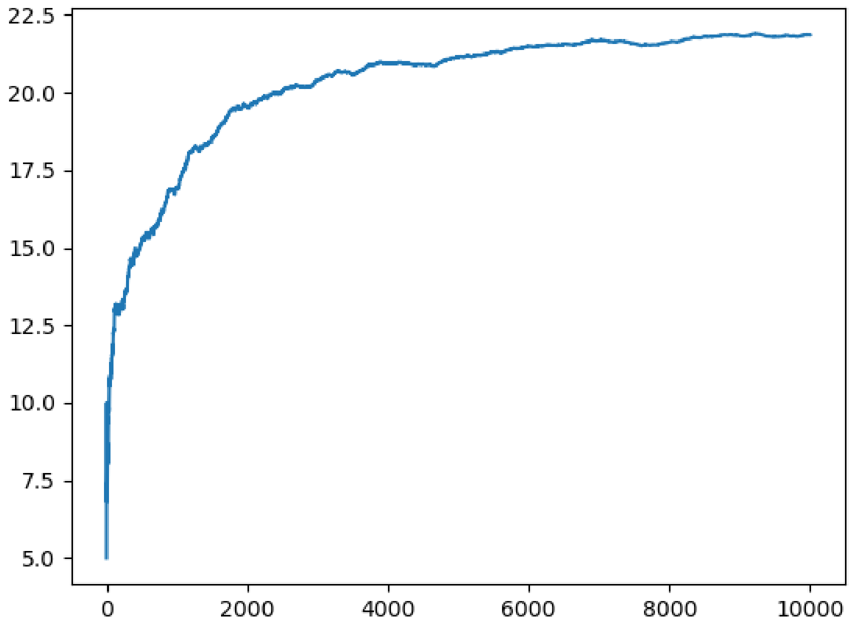
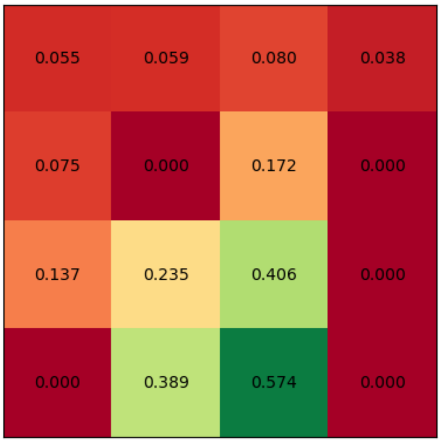


Figure 4: Q-Learning training length



0.055	0.059	0.080	0.038
0.075	0.000	0.172	0.000
0.137	0.235	0.406	0.000
0.000	0.389	0.574	0.000

Figure 5: Q-Learning V

c)

An optimal policy gets calculated. This policy would not be optimal in a slippery environment.

For Sarsa: see Figure 7 & Figure 8 & Figure 9.

↓	→	↓	←
↓	←	↓	←
→	↓	↓	←
←	→	→	←

Table 3: det Sarsa policy

For Q-Learning: see Figure 10 & Figure 11 & Figure 12.

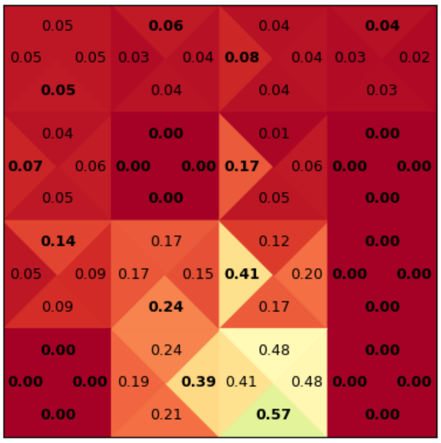


Figure 6: Q-Learning Q

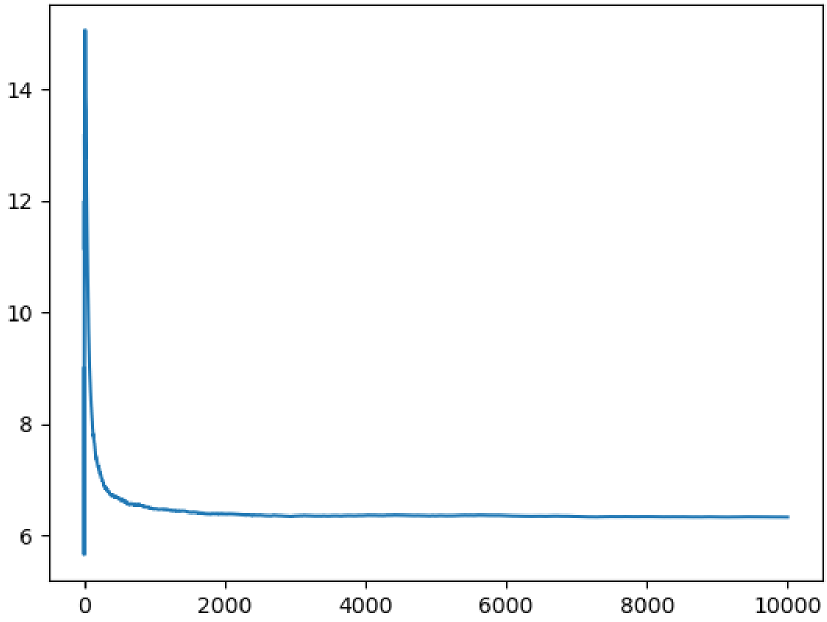


Figure 7: det Sarsa training length

↓	→	↓	←
↓	←	↓	←
→	→	↓	←
←	→	→	←

Table 4: det Q-Learning policy

d)

For Sarsa: see Figure 7 & Figure 8 & Figure 9.

For Q-Learning: see Figure 16 & Figure 17 & Figure 18.

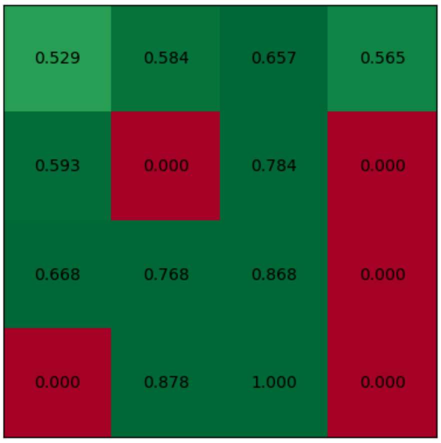


Figure 8: det Sarsa V

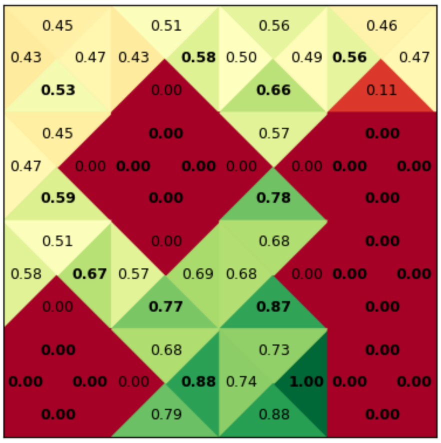


Figure 9: det Sarsa Q

→	→	→	→	→	→	→	←
↑	↑	↑	↑	→	→	→	↓
↑	→	←	←	→	↑	↓	↓
→	↑	↑	↑	←	←	→	↓
→	←	↑	←	→	↓	↑	→
←	←	←	↓	↑	←	←	→
←	←	→	←	←	→	←	→
↓	↑	→	←	→	→	↓	←

Table 5: 8x8 Sarsa policy

↑	→	→	→	→	→	→	→
↑	↑	↑	↑	→	→	→	↓
←	←	←	←	→	↑	→	↓
→	↑	↑	↑	←	←	→	↓
←	→	↑	←	→	↓	↑	→
←	←	←	→	↑	←	←	→
←	←	↓	↓	←	←	←	→
↓	↑	←	←	↓	→	↓	←

Table 6: 8x8 Q-Learning policy

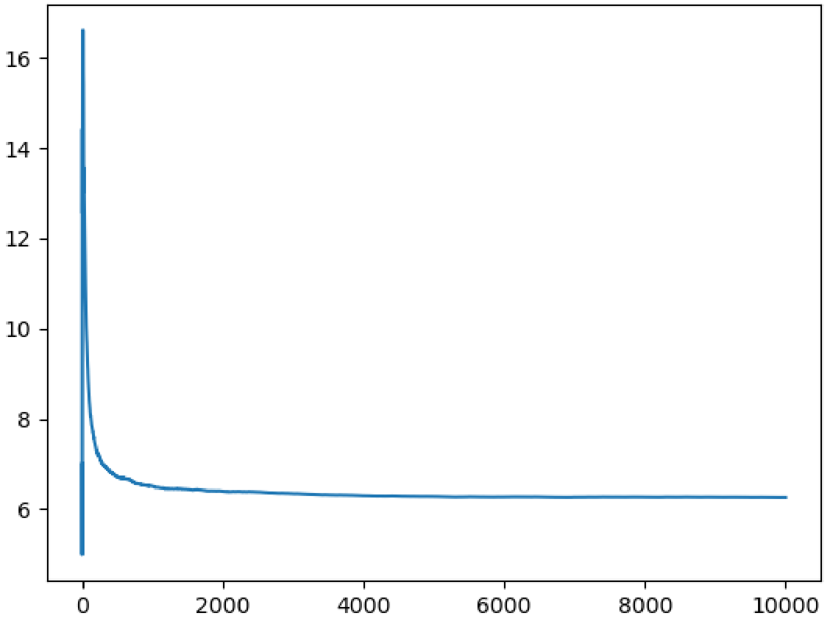
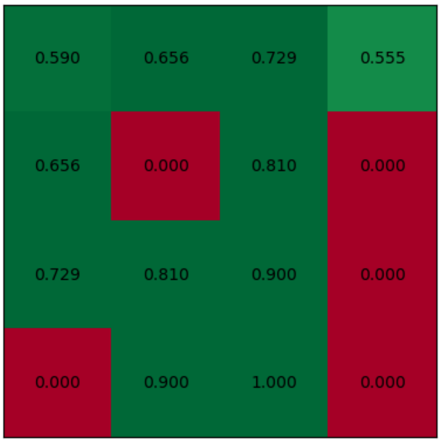


Figure 10: det Q-Learning training length



0.590	0.656	0.729	0.555
0.656	0.000	0.810	0.000
0.729	0.810	0.900	0.000
0.000	0.900	1.000	0.000

Figure 11: det Q-Learning V

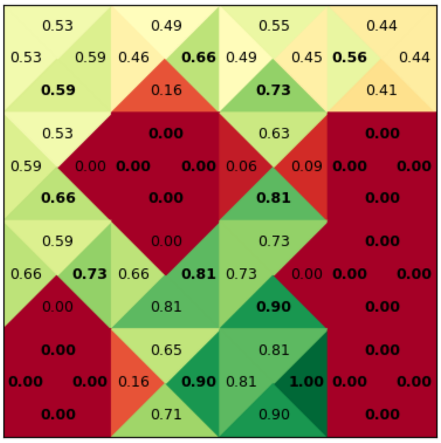


Figure 12: det Q-Learning Q

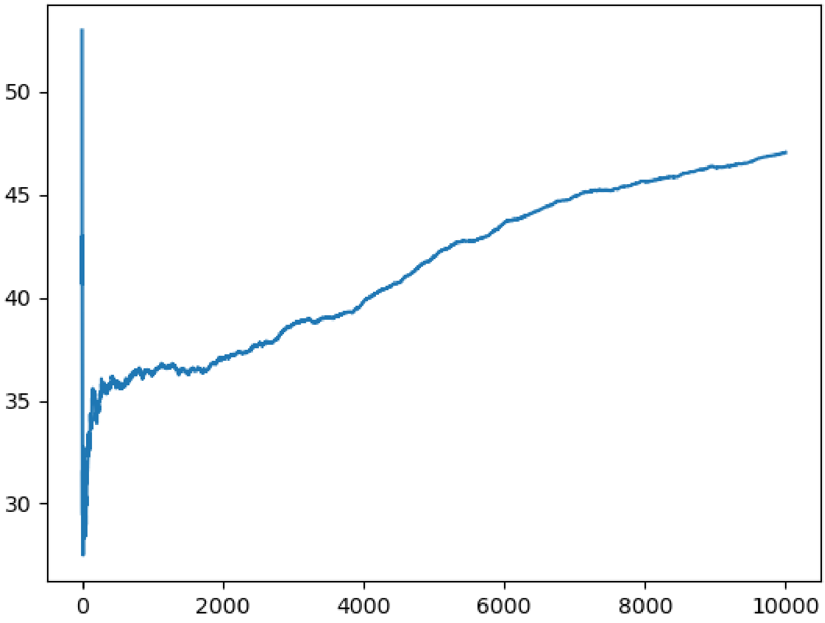


Figure 13: 8x8 Sarsa training length

Felix Bühler - 2973140
Jan Leusmann - 2893121
Jamie Ullerich - 3141241

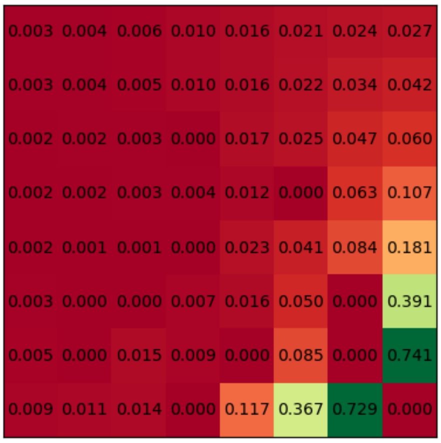


Figure 14: 8x8 Sarsa V

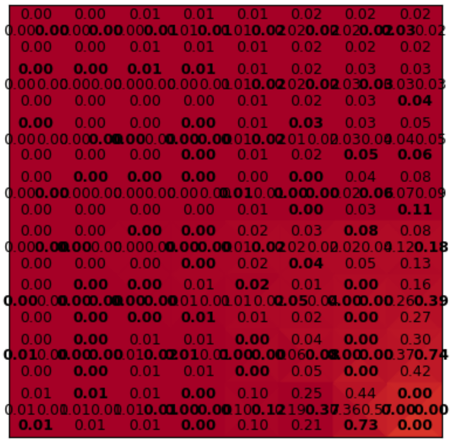


Figure 15: 8x8 Sarsa Q

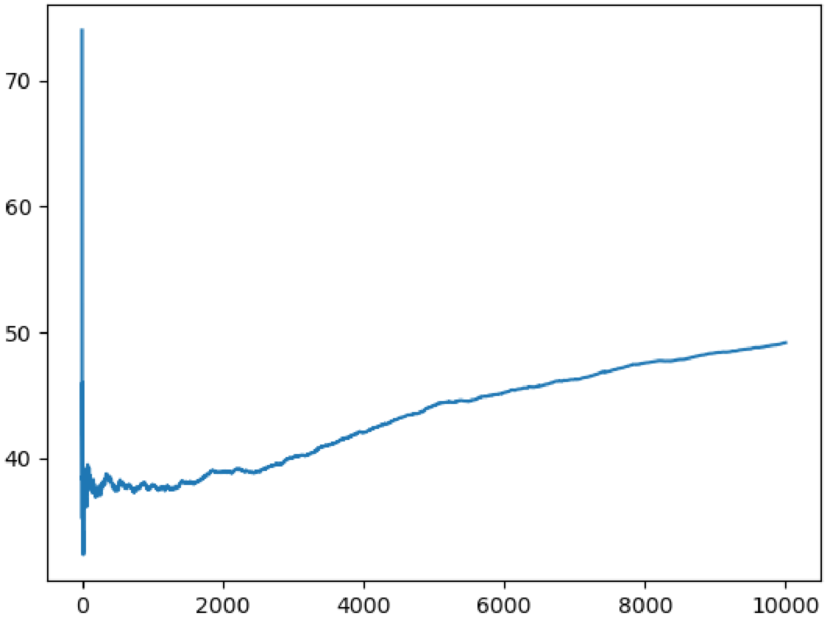
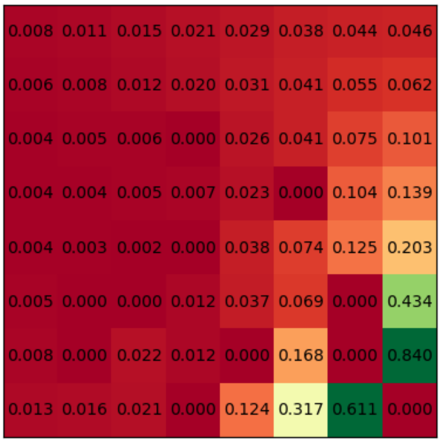


Figure 16: 8x8 Q-Learning training length



0.008	0.011	0.015	0.021	0.029	0.038	0.044	0.046
0.006	0.008	0.012	0.020	0.031	0.041	0.055	0.062
0.004	0.005	0.006	0.000	0.026	0.041	0.075	0.101
0.004	0.004	0.005	0.007	0.023	0.000	0.104	0.139
0.004	0.003	0.002	0.000	0.038	0.074	0.125	0.203
0.005	0.000	0.000	0.012	0.037	0.069	0.000	0.434
0.008	0.000	0.022	0.012	0.000	0.168	0.000	0.840
0.013	0.016	0.021	0.000	0.124	0.317	0.611	0.000

Figure 17: 8x8 Q-Learning V

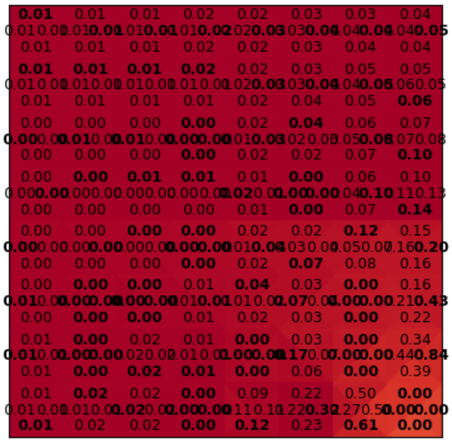


Figure 18: 8x8 Q-Learning Q