EX5

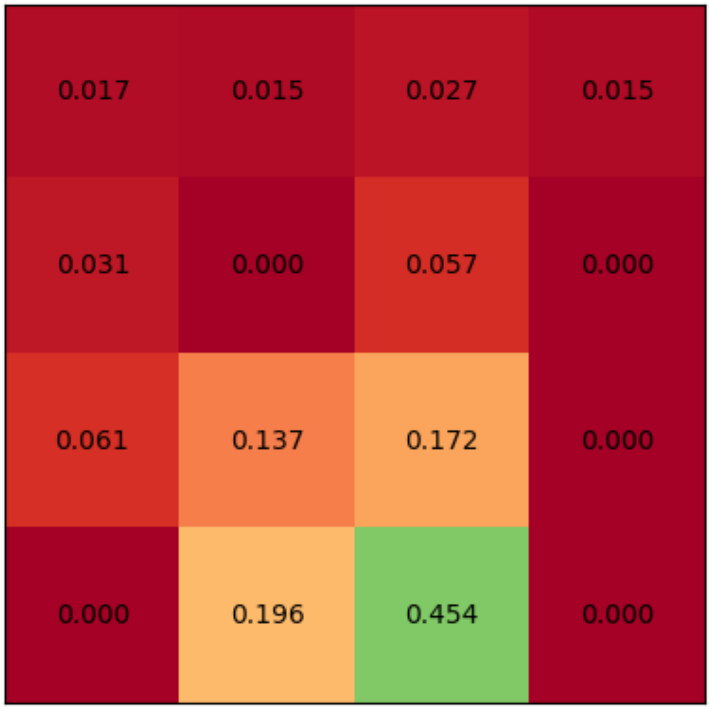
Random Walk

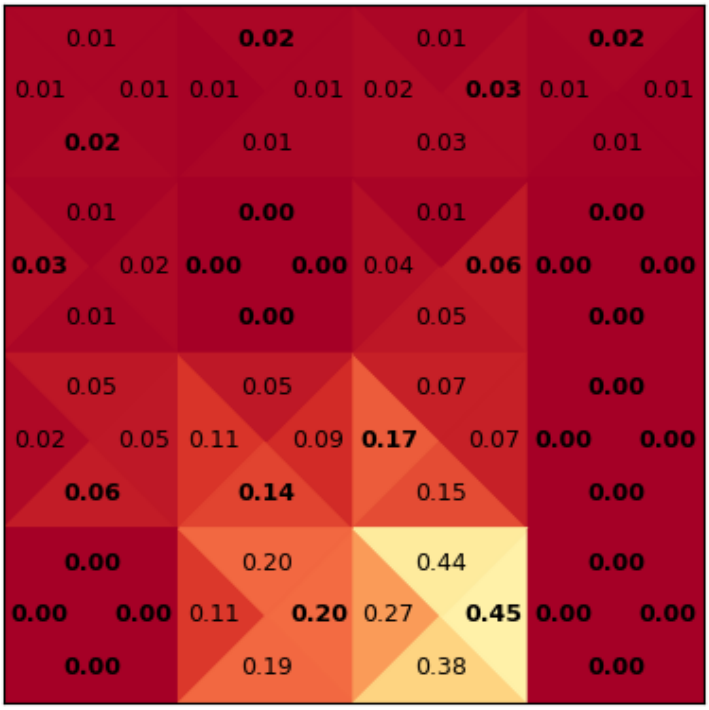
The random walk terminated in the state to the left, what it did before cannot be determined, except we know it did not jump from state E to the right.

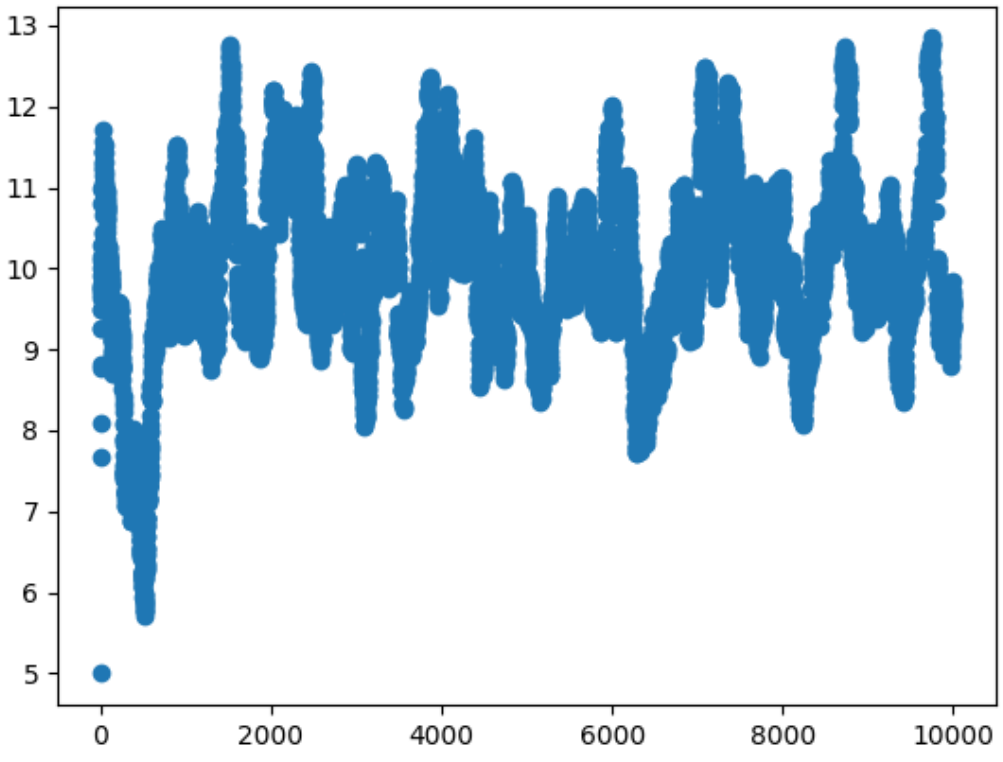
Only the estimate for state A was changed, because with a discount factor of 1, the TD-error was zero for all other states.

$$V(A) \leftarrow V(A) + 0.1\,(0 + 1 \cdot 0 - 0.5)$$
$$= 0.5 - 0.05$$
$$= 0.45$$

Sarsa

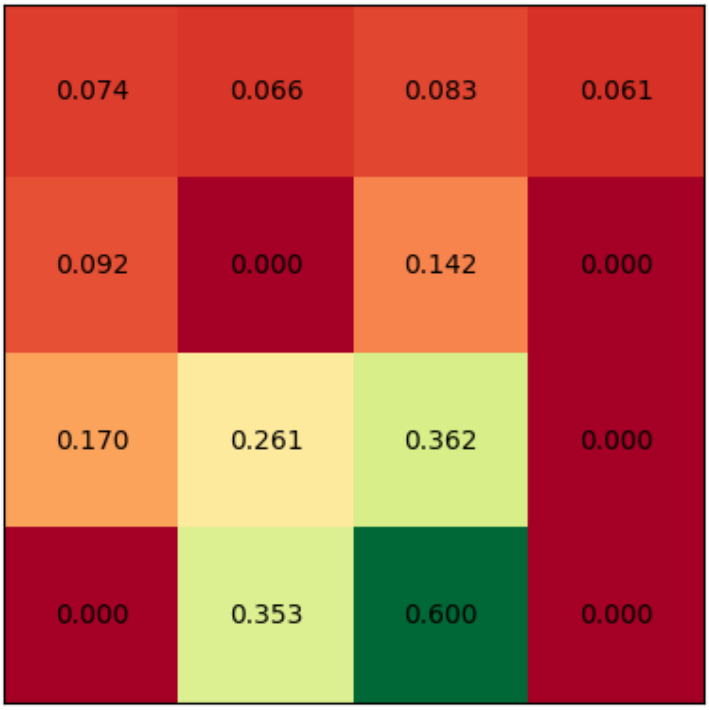



0.01 0.02 0.01 0.02
0.01 0.01 0.01 0.01 0.02 0.03 0.01 0.01
0.02 0.01 0.03 0.01
0.01 0.00 0.01 0.00
0.03 0.02 0.00 0.00 0.04 0.06 0.00 0.00
0.01 0.00 0.05 0.00
0.05 0.05 0.07 0.00
0.02 0.05 0.11 0.09 0.17 0.07 0.00 0.00
0.06 0.14 0.15 0.00
0.00 0.20 0.44 0.00
0.00 0.00 0.11 0.20 0.27 0.45 0.00 0.00
0.00 0.19 0.38 0.00

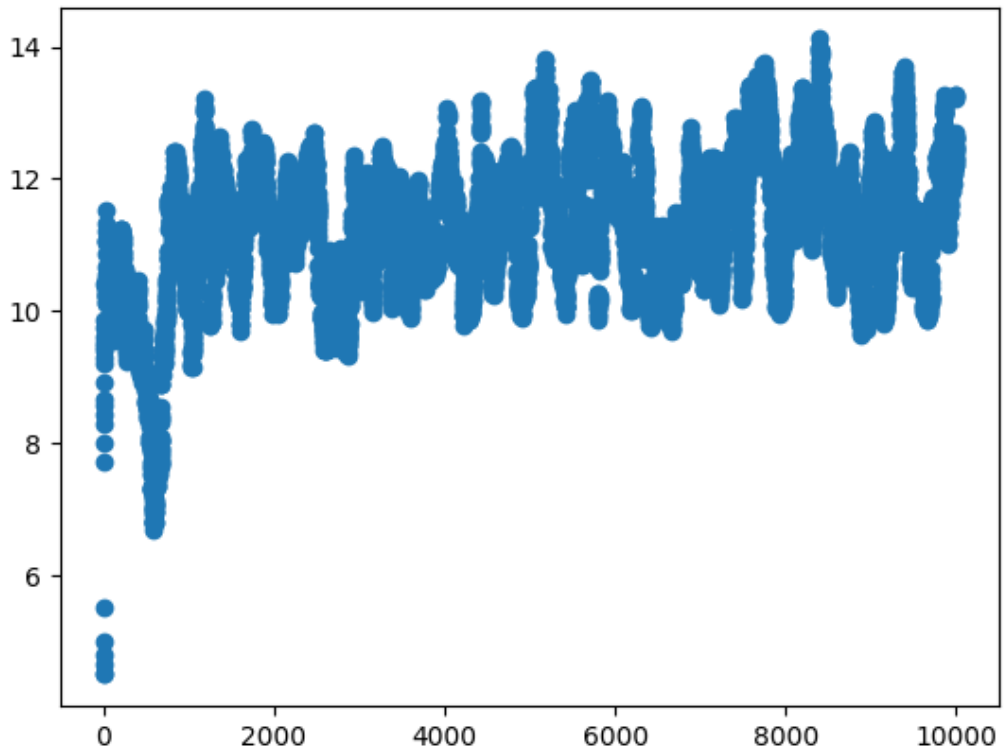

13
12
11
10
9
8
7
6
5
0
2000
4000
6000
8000
10000

The average episode length does not change during training. The total average episode length is 9,997

Q-Learning


0.074
0.066
0.083
0.061
0.092
0.000
0.142
0.000
0.170
0.261
0.362
0.000
0.000
0.353
0.600
0.000


0.06
0.07
0.06
0.06
0.07
0.07
0.04
0.04
0.08
0.08
0.04
0.04
0.07
0.03
0.07
0.04
0.07
0.00
0.03
0.00
0.09
0.07
0.00
0.00
0.14
0.10
0.00
0.00
0.08
0.00
0.09
0.00
0.17
0.11
0.09
0.00
0.06
0.11
0.22
0.17
0.25
0.29
0.00
0.00
0.11
0.26
0.36
0.00
0.00
0.30
0.42
0.00
0.00
0.00
0.21
0.35
0.37
0.50
0.00
0.00
0.00
0.30
0.60
0.00



The average episode length for Q-learning is slightly longer compared to SARSA. The total average is 11,2754

The console output for the small stochastic environment is:

current environment:

SFFF

FHFH

FFFH

HFFG

Running sarsa...

↓ ↑ → ↑

← H → H

↓ ↓ ← H

H → → G

episode_length: 9.997

Running qlearning

← ↑ ← ↑

← H ← H

↑ ↓ ↓ H

H → ↓ G

episode_length: 11.2754

And for the deterministic environment:

current environment:

SFFF

FHFH

FFFH

HFFG

Running sarsa...

↓ → ↓ ←

↓ H ↓ H

→ → ↓ H

H → → G

episode_length: 7.772

Running qlearning

↓ → ↓ ←

↓ H ↓ H

→ ↓ ↓ H

H → → G

episode_length: 7.3249

When changing to the deterministic environment, both algorithms still not create the same policy, but the average episode length drops significantly.

The larger environment exhibits the same behaviour as the small one.