Blatt 4

Q learning

①.

Update rule of Q-learning for updating the Q-function after a transition from a state i state to a state j using action u and observing immediate reward $r(i,u)$:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left(R + \gamma \max_{A'} Q(S',A') - Q(S,A)\right)$$

- $Q(S,A)$ ← : Q-function → new value
- $Q(S,A)$: Q-function → old value
- α : learning factor → "How much does new estimate effect $Q(S,A)$"
- R : Reward
- γ : discounting factor
- $\max_{A'} Q(S',A')$: maximum Q given action
- $Q(S,A)$: Q-function → old value

How would you handle transitions to or within the goal state?

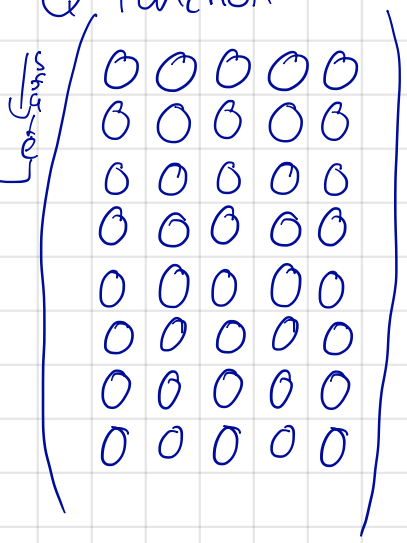
If the target is the goal state just use the reward R as value of the Q-function. Different from the lecture we would use a slightly different update rule: $Q(S,A) \leftarrow (1-\alpha) Q(S,A) \ldots$

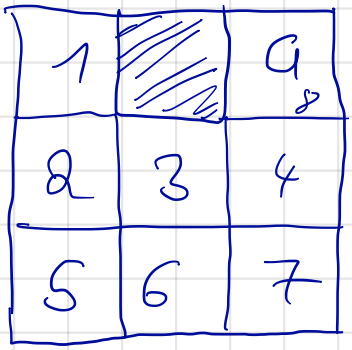
⇒ Then if the target is reached the Q-function value remains the reward R of the goal state

2)

8 states with 5 actions each.

Q-function → actions

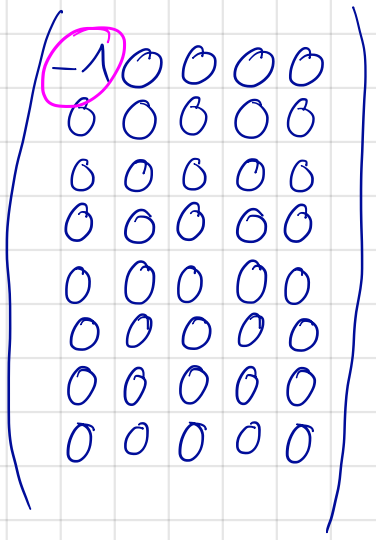


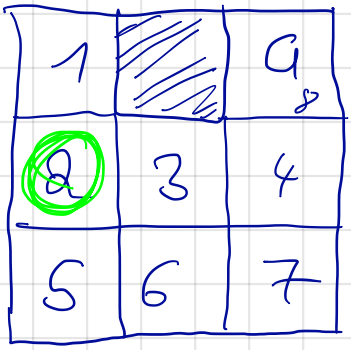


$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(R + \gamma \max_{A'} Q(S',A') - Q(S,A)\right)$$

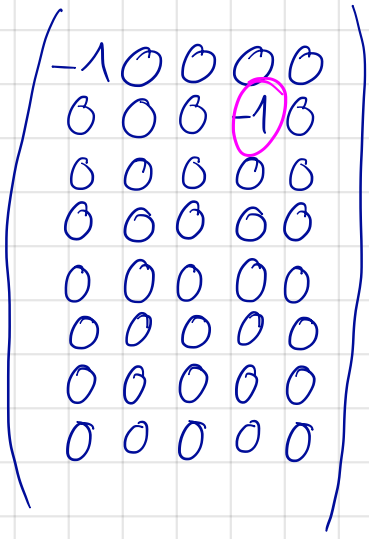
Action: 1 = down
2 = up
3 = left
4 = right
5 = stay

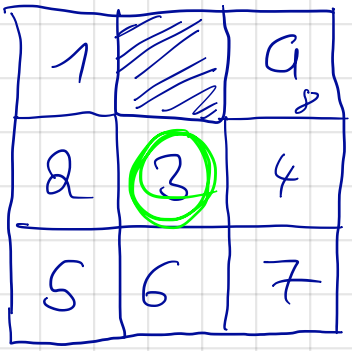
State 1 Action 1
("down")





State 2 Action 4
("right")

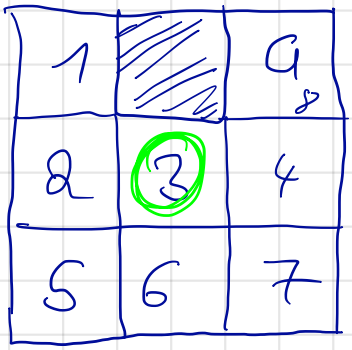




State 3 Action 2
("up")

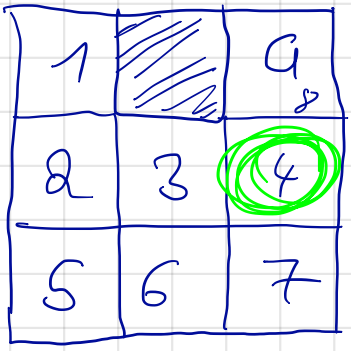
$$\begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

State 3 Action 4

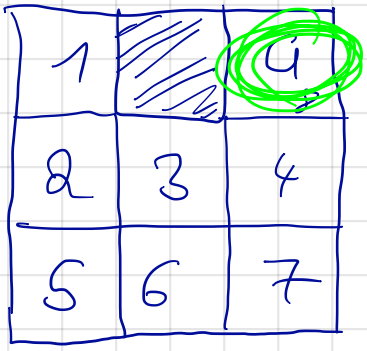


$$\begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

State 4 Action 2



$$\begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



improved Q-function after initial episode