Workshop 10

Policy Iteration

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle$

Output: Policy π

Set V^π to arbitrary value function; e.g., $V^\pi(s) = 0$ for all s.

Set π to arbitrary policy; e.g. $\pi(s) = a$ for all s, where $a \in A$ is an arbitrary action.

Repeat

- Compute $V^\pi(s)$ for all s using policy evaluation

  For each $s \in S$

  $\pi(s) \leftarrow \mathbf{argmax}_{a \in A(s)} Q^\pi(s, a)$

Until π does not change

Step 1: Policy Evaluation

Input: π the policy for evaluation, V^π value function, and MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle$

Output: Value function V^π

Repeat

$\Delta \leftarrow 0$

For each $s \in S$

$$\underbrace{V'^\pi(s) \leftarrow \sum_{s' \in S} P_{\pi(s)}(s' \mid s) \left[r(s, a, s') + \gamma V^\pi(s') \right]}_{\text{Policy evaluation equation}}$$

$\Delta \leftarrow \max(\Delta, |V'^\pi(s) - V^\pi(s)|)$

$V^\pi \leftarrow V'^\pi$

Until $\Delta \leq \theta$

Step 2: Policy Improvement

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle$

Output: Policy π

Set V^π to arbitrary value function; e.g., $V^\pi(s) = 0$ for all s.

Set π to arbitrary policy; e.g. $\pi(s) = a$ for all s, where $a \in A$ is an arbitrary action.

Repeat

Compute $V^\pi(s)$ for all s using policy evaluation

For each $s \in S$

$\pi(s) \leftarrow \mathbf{argmax}_{a \in A(s)} Q^\pi(s, a)$

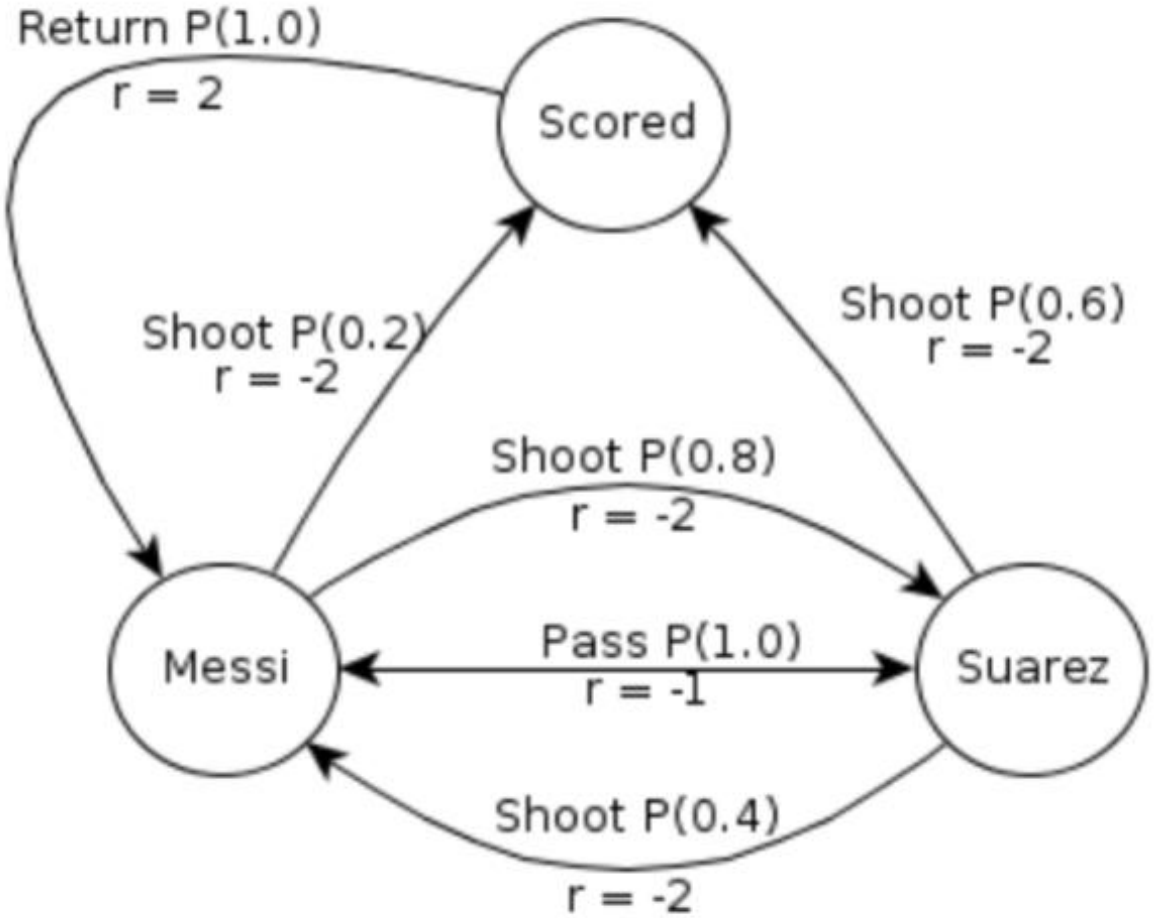
Until π does not change

Problem 1

Consider again the two football-playing robots, Messi and Suarez, from an earlier tutorial. Recall that there are three states: *Messi*, *Suarez* (denoting who has the ball), and *Scored* (denoting that a goal has been scored); and the following action descriptions:

- If Messi shoots, he has 0.2 chance of scoring a goal and a 0.8 chance of the ball going to Suarez. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If Suarez shoots, he has 0.6 chance of scoring a goal and a 0.4 chance of the ball going to Messi. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If either player passes, the ball will reach its intended target with a probability of 1.0. Passing the ball incurs a cost 1 (or a reward of -1).
- If a goal is scored, the only action is to return the ball to Messi, which has a probability of 1.0 and has a reward of 2. Thus the reward for scoring is modelled by giving a reward of 2 when *leaving* the goal state.

Problem 1


Return P(1.0)
r = 2
Scored
Shoot P(0.2)
r = -2
Shoot P(0.6)
r = -2
Shoot P(0.8)
r = -2
Messi
Pass P(1.0)
r = -1
Suarez
Shoot P(0.4)
r = -2

Problem 1

Consider the following policy update table and policy evaluation table, with discount factor $\gamma = 0.8$:

Iter	$Q(Messi, P)$	$Q(Messi, S)$	$Q(Suarez, P)$	$Q(Suarez, S)$	$Q(Scored)$
0	0	0	0	0	0
1					
2	-4.194	-4.772	-4.355	-3.993	-1.355

Apply two iterations of policy iteration. Finish both tables and show the working for the policy evaluation and policy update.

What is the policy after two iterations?

Iter	$\pi(Messi)$	$\pi(Suarez)$	$\pi(Scored)$
0	Pass	Pass	Return
1			Return
2			Return

Problem 1

$$
\begin{aligned}
V^{\pi}(Messi) &= Q^{\pi}(Messi, Pass) \\
&= P_{pass}(Suarez|Messi)[r(Messi, pass, Suarez) + \gamma \cdot V^{\pi}(Suarez)]
\end{aligned}
$$

$$
\begin{aligned}
V^{\pi}(Suarez) &= Q^{\pi}(Suarez, Pass) \\
&= P_{pass}(Messi|Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V^{\pi}(Messi)]
\end{aligned}
$$

$$
\begin{aligned}
V^{\pi}(Scored) &= Q^{\pi}(Scored, return) \\
&= P_{return}(Messi|Scored)[r(Scored, return, Messi) + \gamma \cdot V^{\pi}(Messi)]
\end{aligned}
$$

Problem 1

$$
\begin{aligned}
V^{\pi}(Messi) &= Q^{\pi}(Messi, Pass) \\
&= P_{pass}(Suarez|Messi)[r(Messi, pass, Suarez) + \gamma \cdot V^{\pi}(Suarez)] \\
&= \gamma \cdot V^{\pi}(Suarez) - 1 \\
V^{\pi}(Suarez) &= Q^{\pi}(Suarez, Pass) \\
&= P_{pass}(Messi|Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V^{\pi}(Messi)] \\
&= \gamma \cdot V^{\pi}(Messi) - 1 \\
V^{\pi}(Scored) &= Q^{\pi}(Scored, return) \\
&= P_{return}(Messi|Scored)[r(Scored, return, Messi) + \gamma \cdot V^{\pi}(Messi)] \\
&= \gamma \cdot V^{\pi}(Messi) + 2
\end{aligned}
$$

Problem 1

Then solve a very basic linear algebra about $V^{\pi}(Messi)$ and $V^{\pi}(Suarez)$:

$$V^{\pi}(Messi) = 1/(\gamma - 1)$$
$$V^{\pi}(Suarez) = 1/(\gamma - 1)$$
$$V^{\pi}(Scored) = 3 + 1/(\gamma - 1)$$

Problem 1

Iter	$Q^{\pi}(Messi, P)$	$Q^{\pi}(Messi, S)$	$Q^{\pi}(Suarez, P)$	$Q^{\pi}(Suarez, S)$	$Q^{\pi}(Scored)$
0	0	0	0	0	0
1	-5	-5.52	-5	-4.56	-2
2	-4.194	-4.772	-4.355	-3.993	-1.355

Problem 1

Iter	$\pi(Messi)$	$\pi(Suarez)$	$\pi(Scored)$
0	Pass	Pass	Return
1	Pass	Shoot	Return
2	Pass	Shoot	Return

Q function approximation

- Design for when |S| * |A| is too big, Value Q(s, a) table cannot converge in time
- **Idea**: Convert |S| * |A| pair into a set of features with weights
- Can still do q learning, but approximate q learning
- Learn wight instead of Q(s, a) values

Advantages: can solve relatively large problem

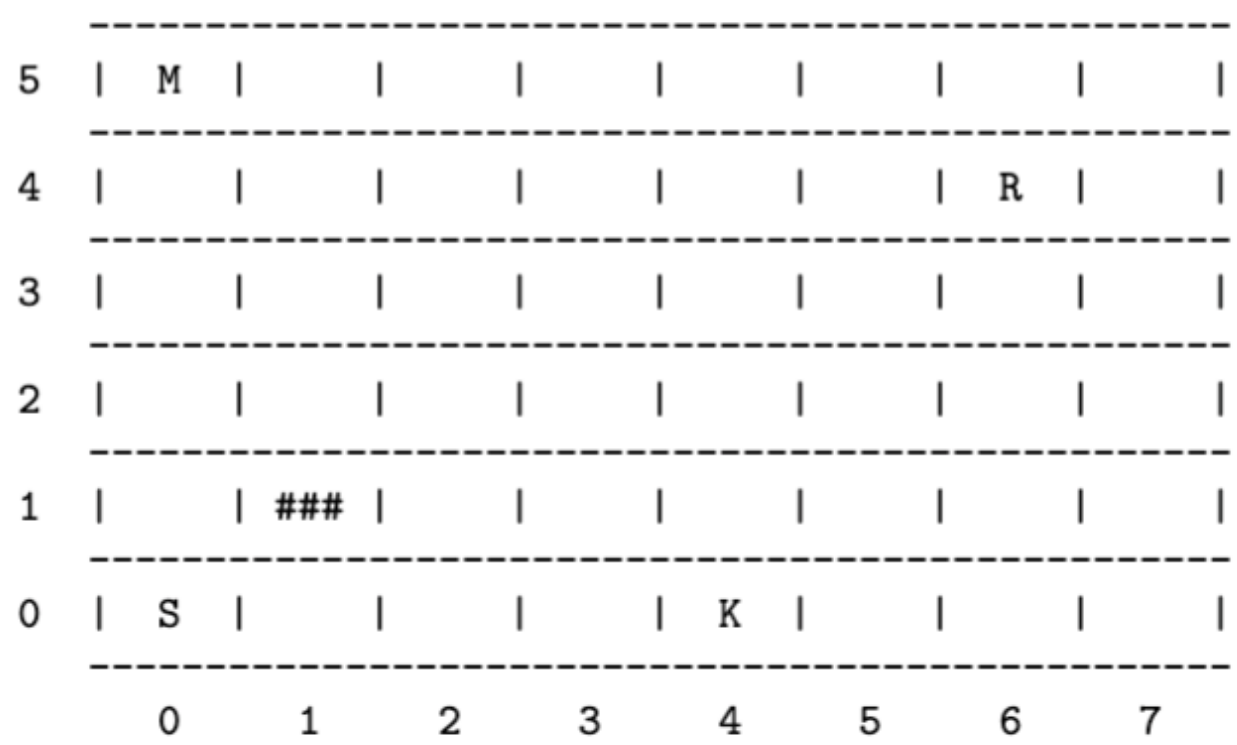
Disadvantages: largely depend on how to implement/design features

Problem 2

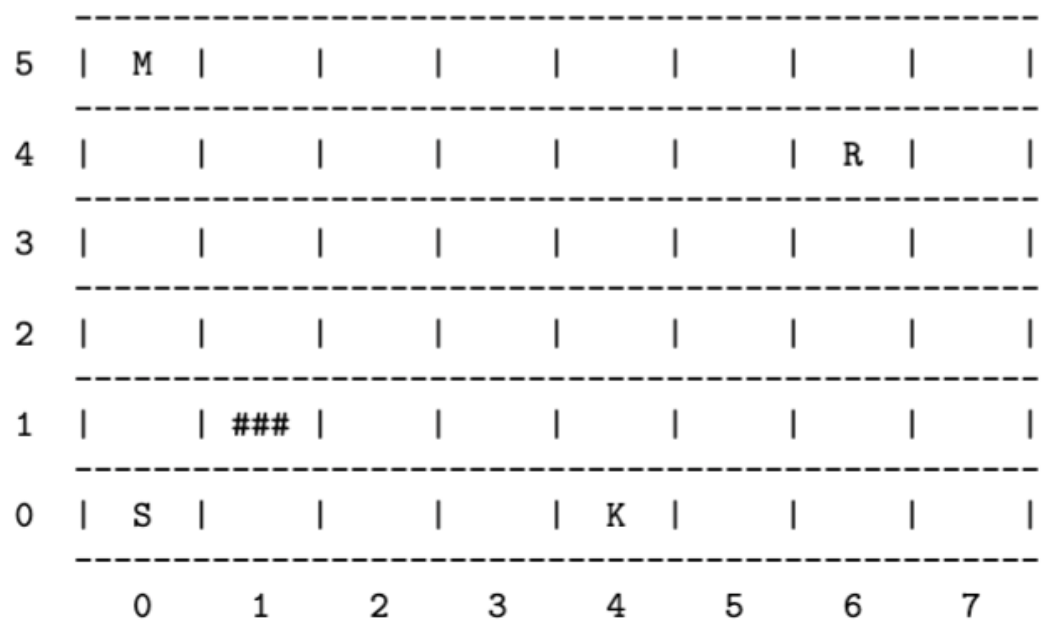
Consider a robotic helper at a hospital that delivers items to staff. The robot is given a task to deliver a treatment kit to a medical specialist in a room. The robot has to pickup the kit from the storeroom, but first has to go to get the key for the storeroom. However, it does not know in advanced whether the key will be there.

The robot will receive a reward of +10 for delivering the kit, and a reward of +5 for going to the room to inform the specialist that the key is missing. There are no other rewards.

Consider this as the following map, where S is the starting state, K is the key rack, M is the medical store room, and R is the room where the store is to be delivered.



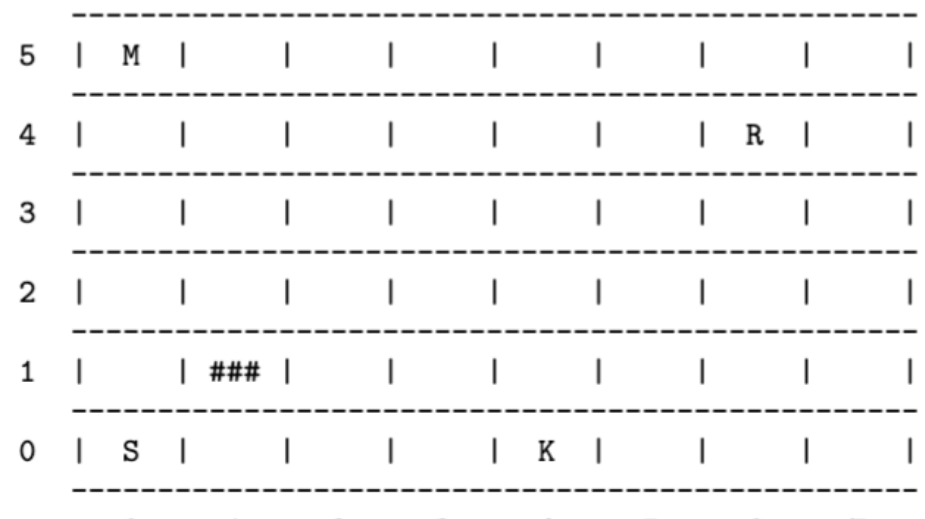
Problem 2



S = {<x, y, K, R> | x belongs to {0…7},
y belongs to {0…5},
K belongs to {0, 1, 2},
R belongs to {0, 1}

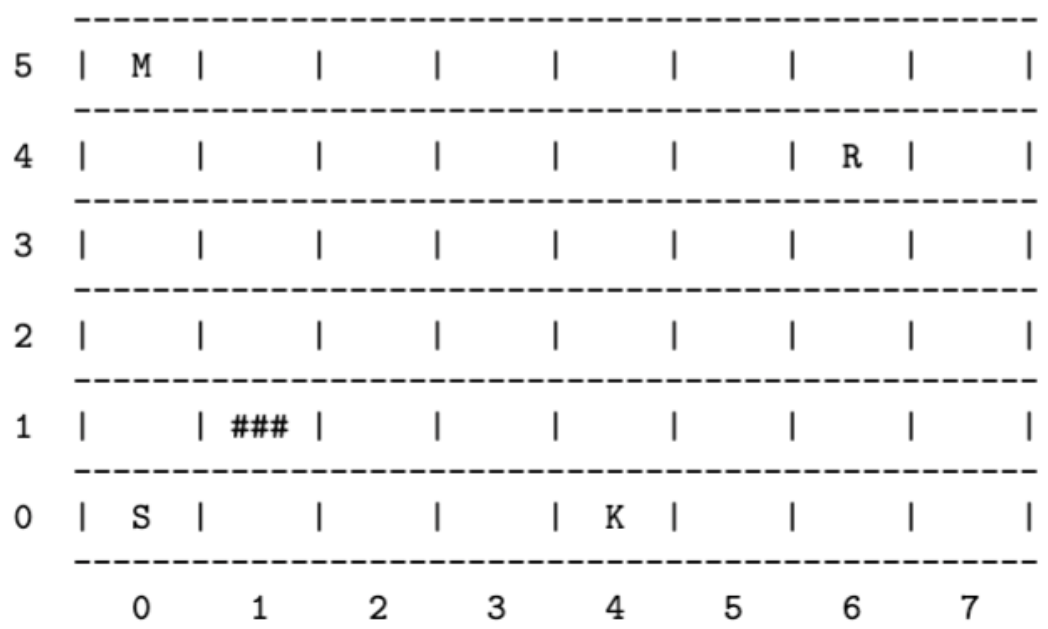
Your task: design a potential function for this problem. You can assume that you can know the the position of the agent, the position of S, M, K, and R, and you can see the a variable Key with values 0, 1, and 2, where 0 indicates there we do not know if the key is in the room, 1 is the key is in the room, and 2 the key is not in the room, and a Boolean variable Med to indicate whether the agent has the medicale kit. Initially, $Key = 0$ and $Med = False$.

Problem 2




```
if Key == 0:
   return 1 - NormalizedManhattan(s, K)
elif Key == 1 and M == False:
   return 1 - NormalizedManhattan(s, M)
elif Key == 1 and M == True:
   return 1 - NormalizedManhattan(s, R)
elif Key == 2:
   return 1 - NormalizedManhattan(s, R)
```

Problem 3



Using your potential function, perform two different reward shaping updates using Q-learning from state K (4,0) and the agent has found the key.

First, perform for the action Up ending in state (4,1).

Then, assume that the *Right* action had been chosen instead of Up, ending in state (5,0).

Compare the two updates to see whether your reward shaping function has worked.

Assume that $Q(s, a) = 0$ for all s and a, $\gamma = 0.9$ and $\alpha = 0.2$.

1-step TD Update

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

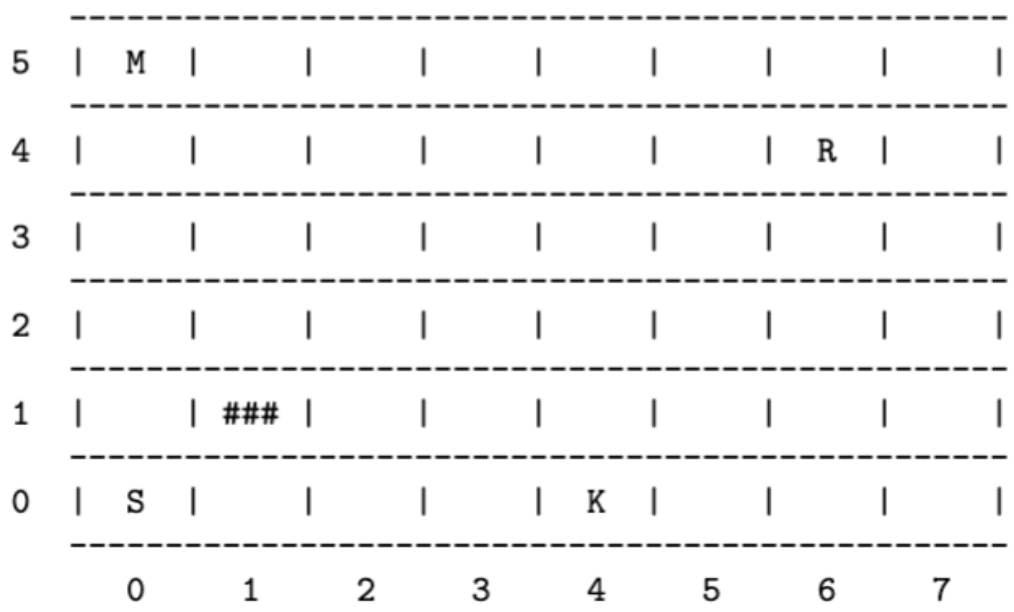
Reward Shaping

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \underbrace{F(s,s')}_{\text{additional reward}} + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

$F : S \times S \to \mathbb{R}$

$$F(s,s') = \gamma\Phi(s') - \Phi(s)$$

Problem 3



Let us s be the current state, s_1 be the state after action Up and s_2 be the state after action $Right$:

$$
\begin{aligned}
s &= ((4,0), Key = 1, M = False) \\
s_1 &= ((4,1), Key = 1, M = False) \\
s_2 &= ((5,0), Key = 1, M = False)
\end{aligned}
$$

$$
\begin{aligned}
\Phi(s) &= 1 - \frac{9}{12} = \frac{3}{12} \\
\Phi(s_1) &= 1 - \frac{8}{12} = \frac{4}{12} \\
\Phi(s_2) &= 1 - \frac{10}{12} = \frac{2}{12}
\end{aligned}
$$

Problem 3

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \underbrace{F(s,s')}_{\text{additional reward}} + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

$F: S \times S \to \mathbb{R}$

$$F(s,s') = \gamma\Phi(s') - \Phi(s)$$

To update the Up action:

$$\begin{aligned} Q(s,Up) &\leftarrow Q(s,Up) + \alpha[r(s,Up,s_1) + F(s,s_1) + \gamma \max Q(s_1,a') - Q(s,Up)] \\ &\leftarrow 0 + 0.2 \times [0 + 0.9 \times \tfrac{4}{12} - \tfrac{3}{12} + 0.9 \times 0 - 0] \\ &\leftarrow 0.01 \end{aligned}$$

$$\begin{aligned} s &= ((4,0), Key = 1, M = False) \\ s_1 &= ((4,1), Key = 1, M = False) \\ s_2 &= ((5,0), Key = 1, M = False) \end{aligned}$$

$$\begin{aligned} \Phi(s) &= 1 - \tfrac{9}{12} = \tfrac{3}{12} \\ \Phi(s_1) &= 1 - \tfrac{8}{12} = \tfrac{4}{12} \\ \Phi(s_2) &= 1 - \tfrac{10}{12} = \tfrac{2}{12} \end{aligned}$$

Problem 3

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \underbrace{F(s,s')}_{\text{additional reward}} + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

$F: S \times S \to \mathbb{R}$

$$F(s,s') = \gamma\Phi(s') - \Phi(s)$$

To update the *Right* action:

$$\begin{aligned} Q(s,Right) &\leftarrow Q(s,Right) + \alpha[r(s,Right,s_1) + F(s,s_2) + \gamma \max Q(s_2,a') - Q(s,Right)] \\ &\leftarrow 0 + 0.2 \times [0 + 0.9 \times \tfrac{2}{12} - \tfrac{3}{12} + 0.9 \times 0 - 0] \\ &\leftarrow -0.02 \end{aligned}$$

$$\begin{aligned} s &= ((4,0), Key = 1, M = False) \\ s_1 &= ((4,1), Key = 1, M = False) \\ s_2 &= ((5,0), Key = 1, M = False) \end{aligned}$$

$$\begin{aligned} \Phi(s) &= 1 - \tfrac{9}{12} = \tfrac{3}{12} \\ \Phi(s_1) &= 1 - \tfrac{8}{12} = \tfrac{4}{12} \\ \Phi(s_2) &= 1 - \tfrac{10}{12} = \tfrac{2}{12} \end{aligned}$$