Week 11: Policy Iteration and Reward Shaping

COMP90054 – AI Planning for Autonomy

Key concepts

- Policy Iteration
- Potential functions and reward shaping

Policy Iteration

Policy Iteration vs Value Iteration?

- Policy Iteration finishes with an optimal policy π after a **finite number of iterations**
- Value Iteration can theoretically require **infinite iterations**

Policy Iteration

Algorithm – Policy Iteration

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

Output: Policy π

Step 1:

Init

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.

Step 2. Repeat

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

$$V^\pi(s) = \sum_{s' \in S} P_{\pi(s)}(s' \mid s)\,[r(s,a,s') + \gamma\, V^\pi(s')]$$

evaluate the V-value/Q-value of all states

Step 3

For each $s \in S$

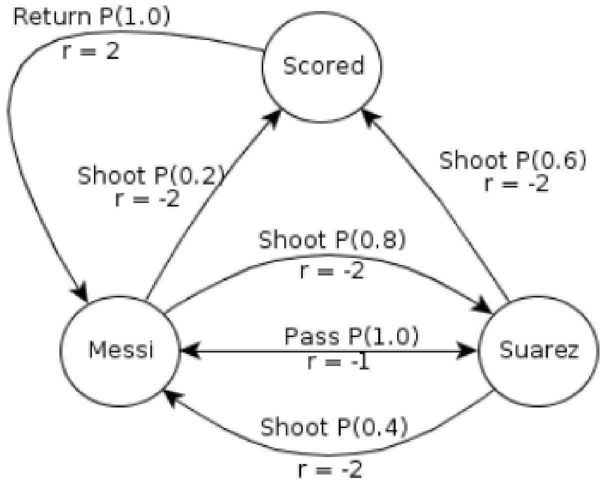
$\pi(s) \leftarrow \text{argmax}_{a \in A(s)} Q^\pi(s,a)$ (policy update)

Until π does not change

Step 4: Check

Problem 1: Policy update

The following diagram shows the transition probabilities and rewards:



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

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	0.0	0.0	0.0
1					
2	-4.194	-4.772	-4.355	-3.993	-1.355

policy evaluation table

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?

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

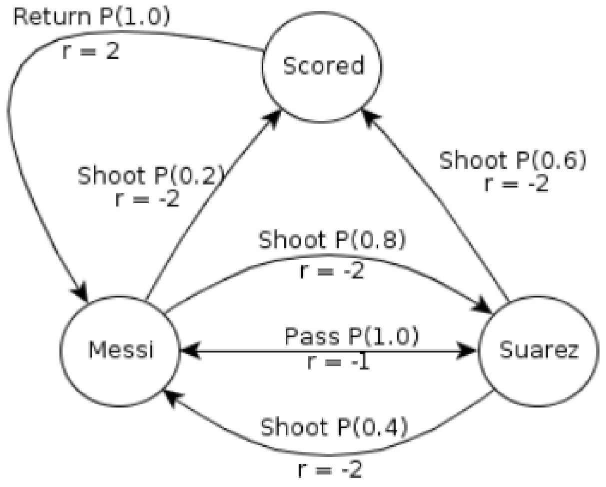
policy update table

Problem 1: Policy update

Step 1: Start with a random policy

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

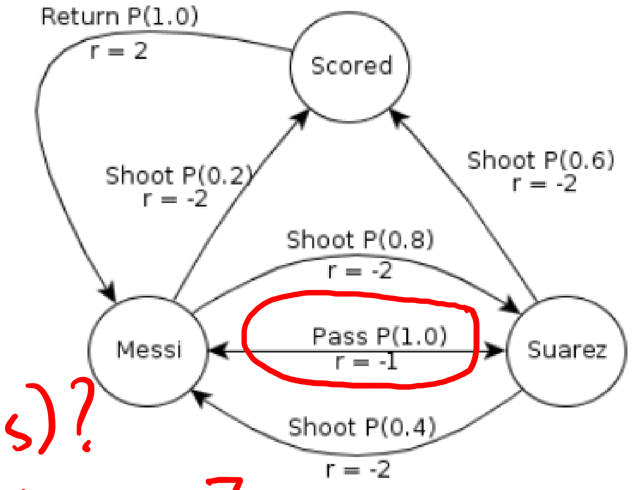
The following diagram shows the transition probabilities and rewards:



Problem 1: Policy update

$$\begin{cases} V^{\pi}(\text{Messi}) = a \\ V^{\pi}(\text{Suarez}) = b \\ V^{\pi}(\text{Scored}) = c \end{cases}$$

The following diagram shows the transition probabilities and rewards:



Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

Iteration 1 - Step 2: Policy Evaluation

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

- $V^{\pi}(\text{Messi}) = Q^{\pi}(\text{Messi, shoot})$ or $Q^{\pi}(\text{Messi, pass})$?

$$\underbrace{V^{\pi}(\text{Messi})}_{a} = Q^{\pi}(\text{Messi, pass}) = \underbrace{P_{pass}(\text{Suarez}|M)}_{1} \times \underbrace{[r + \gamma V^{\pi}(\text{Suarez})]}_{[-1 + 0.8b]}$$

$$\Rightarrow \boxed{a = -1 + 0.8b} \quad \text{Eq 1}$$

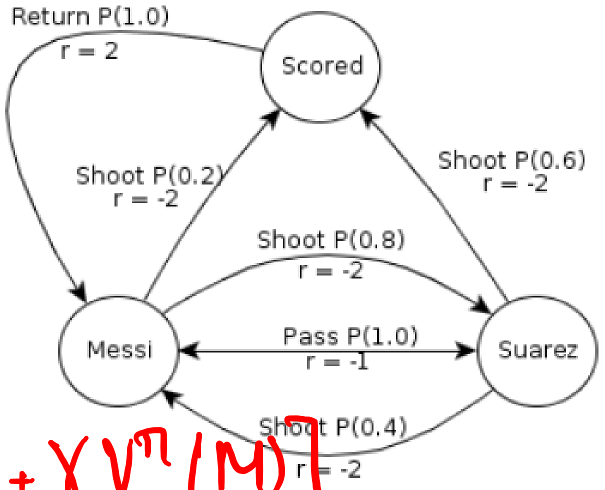
- $V^{\pi}(\text{Suarez}) = Q^{\pi}(\text{Suarez, pass}) = P_{pass}(M|\text{Suarez})[r + \gamma V^{\pi}(M)]$

$$b = 1 \times [-1 + 0.8a]$$

$$\Rightarrow \boxed{b = -1 + 0.8a} \quad \text{Eq 2}$$

Problem 1: Policy update

The following diagram shows the transition probabilities and rewards:



Iteration 1 - Step 2: Policy Evaluation

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

- $V^{\pi}(\text{Scored}) = Q^{\pi}(\text{Scored}, \text{return}) = P_{\text{return}}(M|\text{Scored})[r + \gamma V^{\pi}(M)]$

$$c \quad = 1 \times [2 + 0.8a]$$

$$\Rightarrow \boxed{c = 2 + 0.8a} \quad \text{Eq3}$$

$$\begin{cases} a = -1 + 0.8b \\ b = -1 + 0.8a \\ c = 2 + 0.8a \end{cases} \Rightarrow \begin{cases} a = -5 = V^{\pi}(M) = Q^{\pi}(M, \text{pass}) \\ b = -5 = V^{\pi}(\text{Suarez}) = Q^{\pi}(S, \text{pass}) \\ c = -2 = V^{\pi}(\text{Scored}) \\ \quad = Q^{\pi}(\text{Scored}, \text{return}) \end{cases}$$

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	0.0	0.0	0.0
1	-5		-5		-2
2	-4.194	-4.772	-4.355	-3.993	-1.355

Problem 1: Policy update

Iteration 1 - Step 2: Policy Evaluation

$$\begin{cases} V^\pi(M) = -5 \\ V^\pi(Suarez) = -5 \\ V^\pi(Scored) = -2 \end{cases}$$

$$V^\pi(s) = Q^\pi(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^\pi(s')] \leftarrow$$

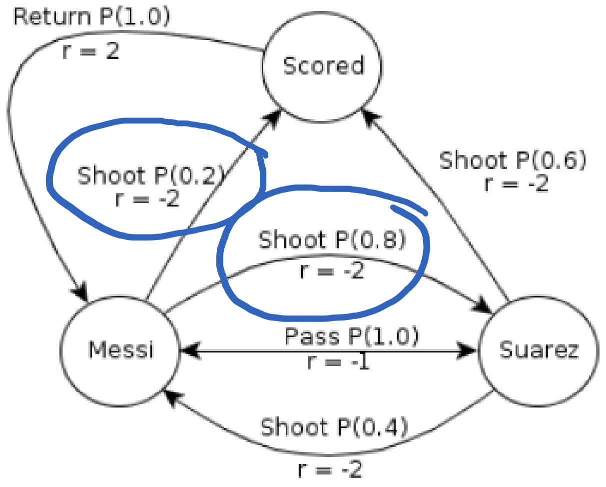
- $Q^\pi(M, shoot) = P_{shoot}(Suarez|M)[r + \gamma V^\pi(Suarez)] + P_{shoot}(Scored|M)[r + \gamma V^\pi(Scored)]$

$$= 0.8[-2 + 0.8 \times (-5)] + 0.2[-2 + 0.8 \times (-2)] = -5.52$$

- $Q^\pi(Suarez, shoot) = \ldots\ldots$

$$= -4.56$$

The following diagram shows the transition probabilities and rewards:



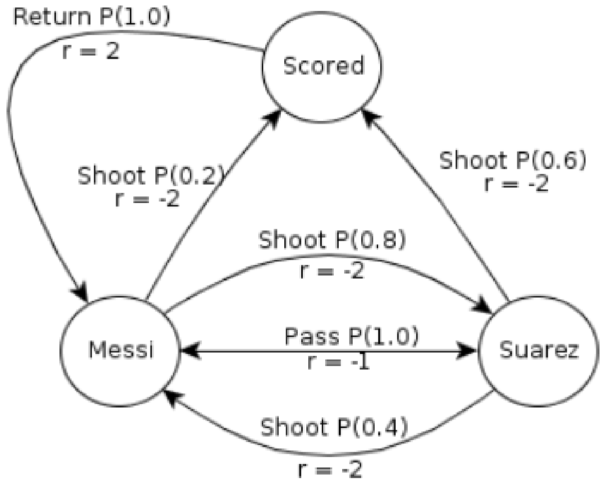
Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	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: Policy update

Iteration 1 - Step 3: Policy Update

The following diagram shows the transition probabilities and rewards:



Update →

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2			Return

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	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

max

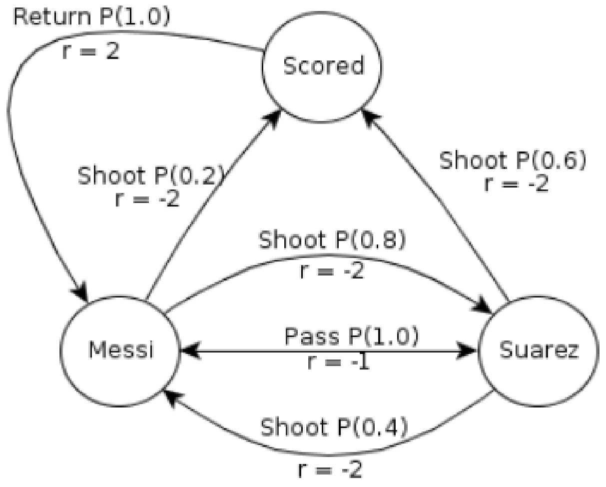
max

Problem 1: Policy update

Iteration 1 - Step 4: When to stop the iteration?

- If the policy changes in Step 3, continue the iteration in Step 2
- If the policy does not change, stop the iteration

The following diagram shows the transition probabilities and rewards:



Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2			Return

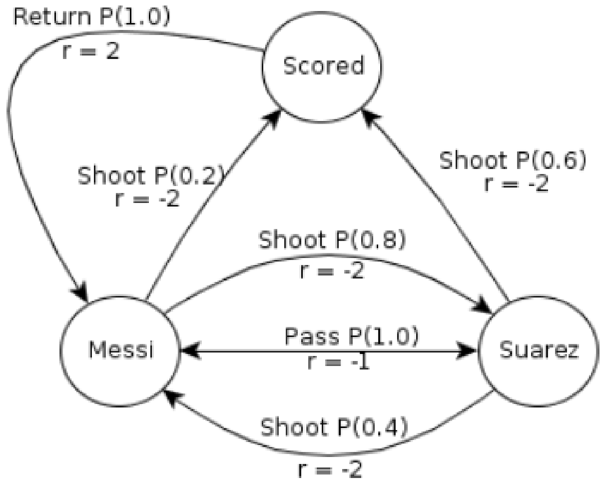
can't stop the iteration ← not the same ←

Problem 1: Policy update

Iteration 2 - Step 2: Policy Evaluation

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

The following diagram shows the transition probabilities and rewards:



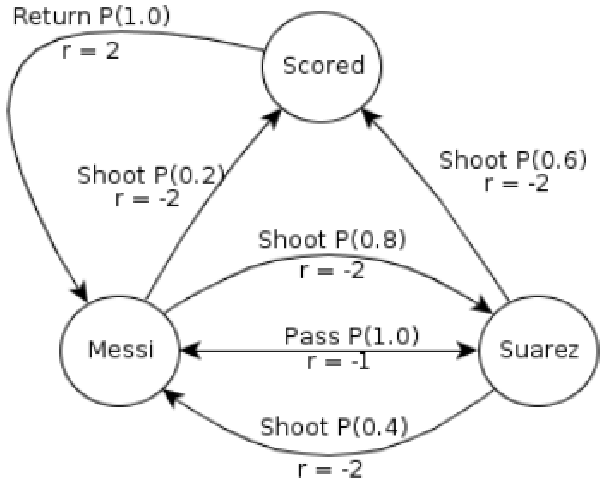
Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2			Return

Problem 1: Policy update

Iteration 2 - Step 2: Policy Evaluation

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

The following diagram shows the transition probabilities and rewards:



Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2			Return

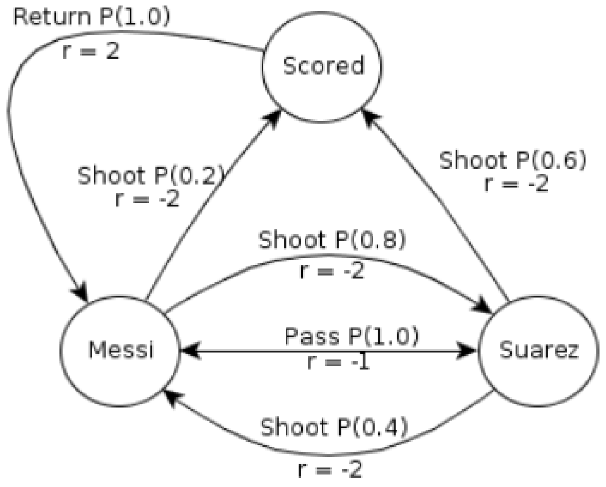
Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	0.0	0.0	0.0
1	−5	−5.52	−5	−4.56	−2
2	-4.194	-4.772	-4.355	-3.993	-1.355

Problem 1: Policy update

Iteration 2 - Step 2: Policy Evaluation

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

The following diagram shows the transition probabilities and rewards:



stop ← same ←

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2	pass	shoot	Return

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	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

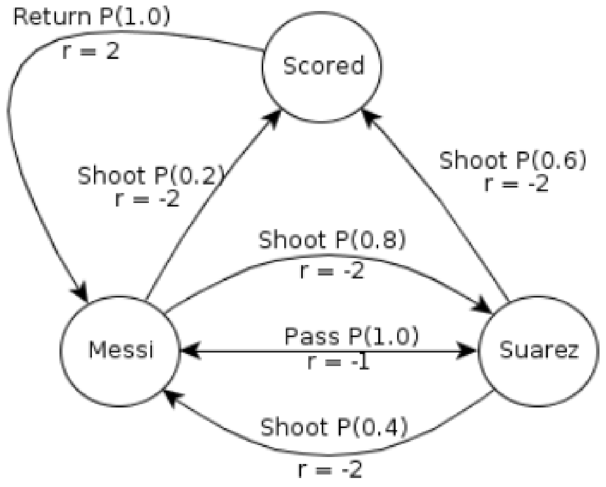
↓ max ↓ max

Problem 1: Policy update

Iteration 2 - Step 3: Policy Update

$$V^{\pi}(s) = Q^{\pi}(s,a) = \sum_{s' \in S} P_{\pi(s)}(s'|s)[r(s,a,s') + \gamma V^{\pi}(s')]$$

The following diagram shows the transition probabilities and rewards:



Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	pass	shoot	Return
2			Return

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suarez, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	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 2: Potential-based Reward shaping

	0	0	0	0	+1 ← G
	0	0	0	0	0
	0	0	0	0	0
S →	0	0	0	0	0

Why do we need **reward shaping**?

- Rewards are sometimes **sparse** (having many zero rewards) -> RL will behave randomly
- Reward shaping is a method in which we can modify the reward function to reward the action that moves us closer to the goal

Potential-based reward shaping function F

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

where Φ is the potential function and $\Phi(s)$ is the potential of state s

An example of the potential function for GridWorld

$$\Phi(s) = 1 - \frac{|x(g) - x(s)| + |y(g) - y(s)|}{width + height - 2}$$

manhattan distance between the current state to the goal state

If the agent moves closer to the goal, $\phi(s) \uparrow$

Problem 2: Potential functions

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.

```
   -------------------------------------------------
5  |  M  |     |     |     |     |     |     |     |
   -------------------------------------------------
4  |     |     |     |     |     |     |  R  |     |
   -------------------------------------------------
3  |     |     |     |     |     |     |     |     |
   -------------------------------------------------
2  |     |     |     |     |     |     |     |     |
   -------------------------------------------------
1  |     | ### |     |     |     |     |     |     |
   -------------------------------------------------
0  |  S  |     |     |     |  K  |     |     |     |
   -------------------------------------------------
```

S= starting pos, K= keyroom, M= med kit, R= staff room (final room)

1) start → get the key → get the med kit → deliver the kit to the staff

S → K → M → R

2) start → can't find the key → go to the staff room to notify the staff

S → K → R

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 agent is holding the key, and 2 the key is not in the room, and a Boolean variable Med to indicate whether the agent has the medical kit. Initially, Key = 0 and Med = False.

Define the potential functions for all cases

Problem 2: Potential functions

```
   -------------------------------------------------
 5 |  M  |     |     |     |     |     |     |     |
   -------------------------------------------------
 4 |     |     |     |     |     |     |  R  |     |
   -------------------------------------------------
 3 |     |     |     |     |     |     |     |     |
   -------------------------------------------------
 2 |     |     |     |     |     |     |     |     |
   -------------------------------------------------
 1 |     | ### |     |     |     |     |     |     |
   -------------------------------------------------
 0 |  S  |     |     |     |  K  |     |     |     |
   -------------------------------------------------
```

K = 0 : the agent hasn't visited the key room
k = 1 : the agent has found the key
k = 2 : the agent can't find the key

M = False : the agent hasn't collected the med kit

M = True : the agent has the med kit

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 agent is holding the key, and 2 the key is not in the room, and a Boolean variable Med to indicate whether the agent has the medical kit. Initially, Key = 0 and Med = False.

current state

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

$(s \rightarrow K)$

$(s \rightarrow M)$

$(s \rightarrow R)$

$(s \rightarrow R)$

Problem 3: Reward shaping update

K → what is the next goal?

K → M

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, γ = 0.9 and α = 0.2.

```
 ---------------------------------------------------------------
5 |  M  |     |     |     |     |     |     |     |
 ---------------------------------------------------------------
4 |     |     |     |     |     |     |  R  |     |
 ---------------------------------------------------------------
3 |     |     |     |     |     |     |     |     |
 ---------------------------------------------------------------
2 |     |     |     |     |     |     |     |     |
 ---------------------------------------------------------------
1 |     | ### |     |     |     |     |     |     |
 ---------------------------------------------------------------
0 |  S  |     |     |     |  K  |     |     |     |
 ---------------------------------------------------------------
     0     1     2     3     4     5     6     7
```

$$\Phi(s) = 1 - \frac{|x(g) - x(s)| + |y(g) - y(s)|}{width + height - 2}$$

W = 8

H = 6

(4,0) s_0

Up → (4,1) s_1

Right → (5,0) s_2

$s_0 = ((4,0), K=1, M=\text{False})$ → goal M(0,5)

$s_1 = ((4,1), K=1, M=\text{False})$ → goal M(0,5)

$s_2 = ((5,0), K=1, M=\text{False})$ → goal M(0,5)

$$\Phi(s_0) = 1 - \frac{|0-4| + |5-0|}{8+6-2} = \frac{3}{12}$$

$$\Phi(s_1) = \frac{4}{12}, \quad \Phi(s_2) = \frac{2}{12}$$

Problem 3: Reward shaping update

Compare the two updates to see whether your reward shaping function has worked.
Assume that Q(s,a) = 0 for all s and a, γ = 0.9 and α = 0.2.

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

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

5	M							
4							R	
3			0	0	0			
2			0	0	0	0	0	
1		###	0	0	0	0	0	
0	S		0	0	K	0	0	

Up $(4,0)\, s_0$ Right

$(4,1)\, s_1$ $(5,0)\, s_2$

$\phi(s_0) = \frac{3}{12}$, $\phi(s_1) = \frac{4}{12}$, $\phi(s_2) = \frac{2}{12}$

1) $s_0 \xrightarrow{\text{Up}} s_1$: $Q(s_0, \text{Up}) = ?$

$F(s_0, s_1) = \gamma\phi(s_1) - \phi(s_0) = 0.9\left(\frac{4}{12}\right) - \frac{3}{12} = 0.05$

$Q(s_0, \text{Up}) = Q(s_0, \text{Up}) + \alpha\left[r + F(s_0, s_1) + \gamma \max_{a'} Q(s_1, a') - Q(s_0, \text{Up})\right]$

$= 0 + 0.2\left[0 + 0.05 + 0.9 \times (0) - 0\right]$

$= 0.01$ → max → Choose Up

2) $s_0 \xrightarrow{\text{Right}} s_2$: $F(s_0, s_2) = \gamma\phi(s_2) - \phi(s_0) = 0.9\left(\frac{2}{12}\right) - \frac{3}{12} = -0.1$

$Q(s_0, \text{Right}) = Q(s_0, \text{Right}) + \alpha\left[r + F(s_0, s_2) + \gamma \max_{a'} Q(s_2, a') - Q(s_0, \text{Right})\right]$

$= 0 + 0.2\left[0 + (-0.1) + 0.9 \times (0) - 0\right] = -0.02$