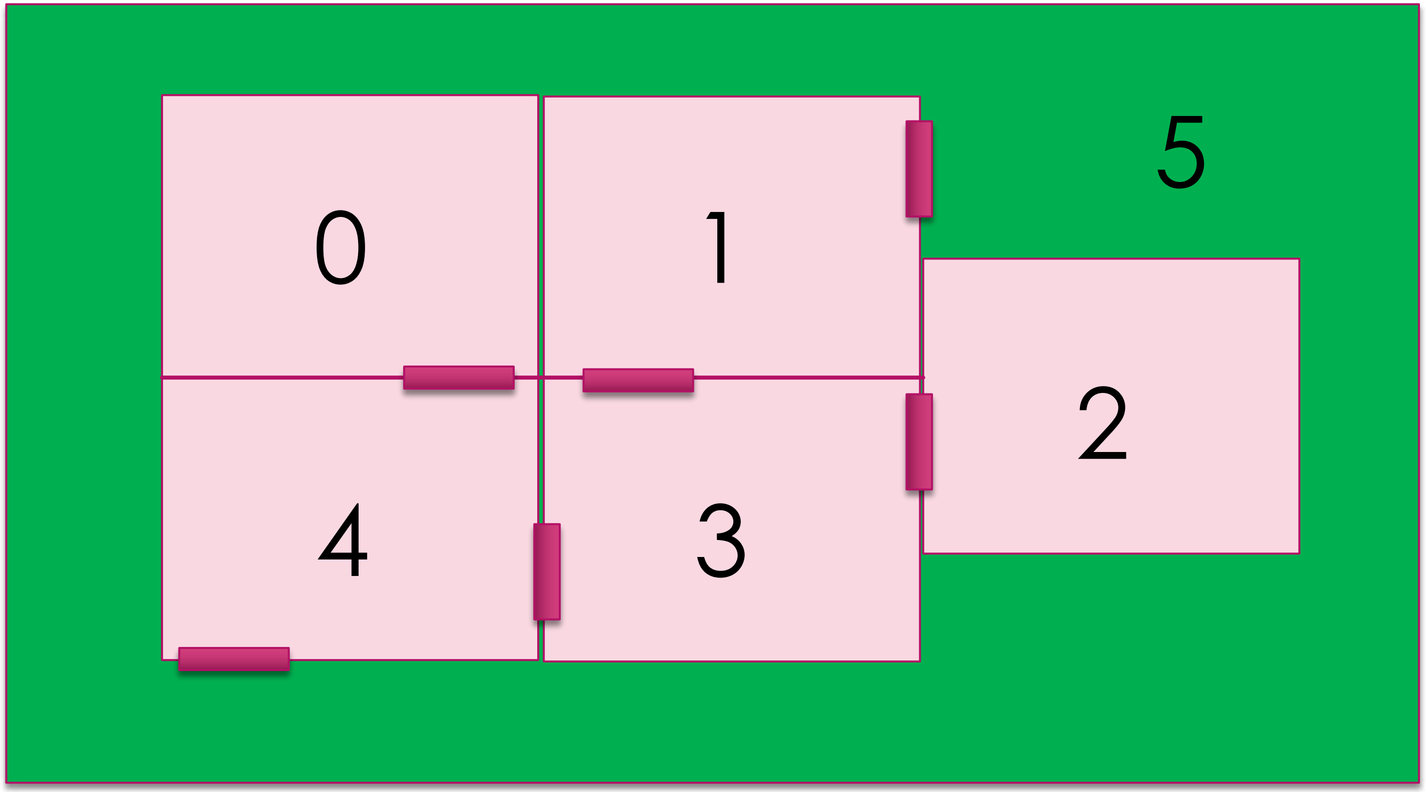
Aprendizado por Reforço: conceitos, aplicações e desafios

ANNA HELENA REALI COSTA
UNIVERSIDADE DE SÃO PAULO
anna.reali@usp.br

Parte 2

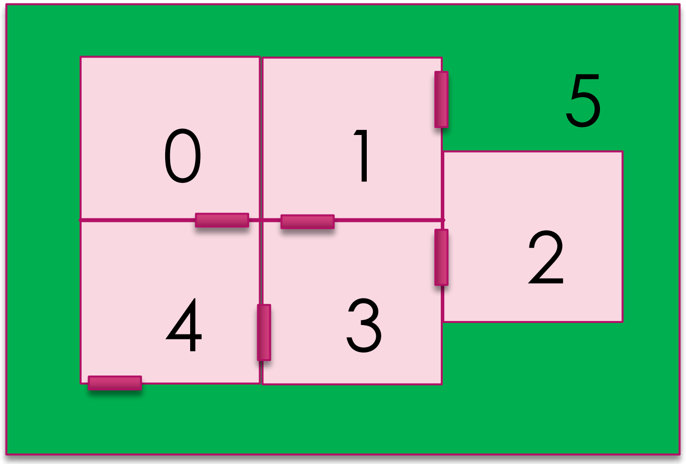
A Simple Example of Q-Learning



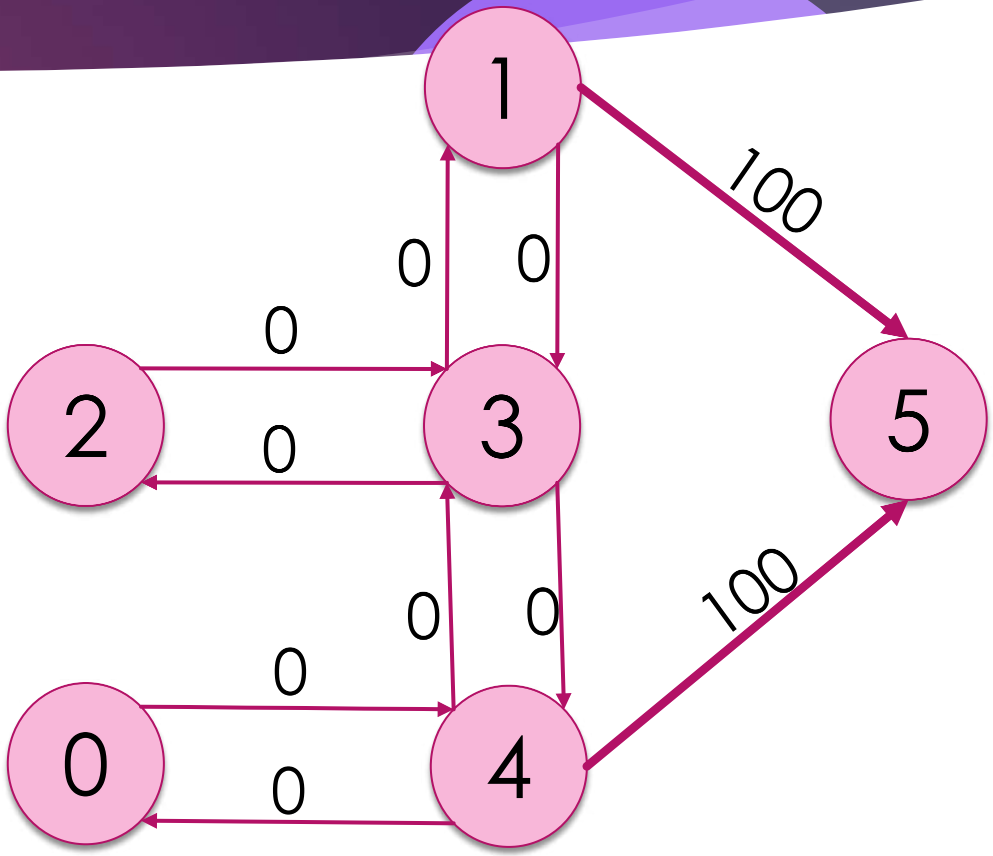
Deterministic world

Goal: to go outside the building

A Simple Example of Q-Learning



$\equiv$



S = {0, 1, 2, 3, 4, 5}

A = A(0)∪A(1)∪A(2)∪A(3)∪A(4)∪A(5)

T: P(s' | s,a) = 1 ∀s, s'∈S, ∀a∈A(s)

R: as indicated in the graph

$\gamma = 0.8$; $\alpha = 1$ $\boldsymbol{Q(s,a)_{t+1} = r(s,a,s') + \gamma \max_{a'} Q(s',a')}$

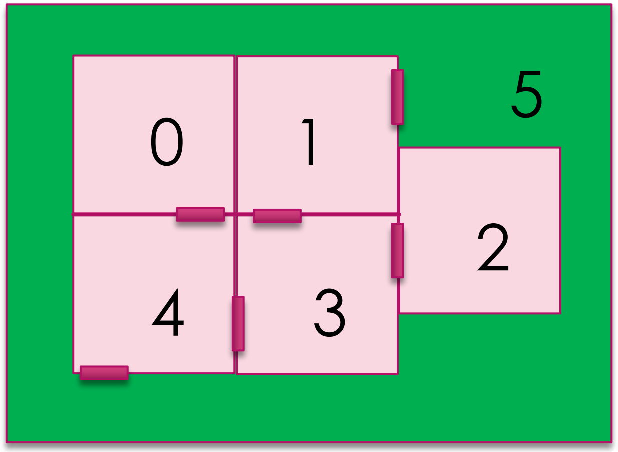
A Simple Example of Q-Learning

Suppose: **initial state = 1**
A(1) = {go to 3, go to 5}

$$Q(s,a)_{t+1} = r(s,a,s') + 0.8 \max_{a'} Q(s',a')$$

Initial Q-table

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0



A Simple Example of Q-Learning

Suppose: **initial state = 1**
A(1) = {go to 3, go to 5}

By random selection: **a = go to 5**
Experience = **(1, go to 5, 5, 100)**

Initial Q-table

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

$$Q(1,5) = 100 + 0.8 \max(Q(5,.))$$
$$= 100 + 0.8 * 0 = 100$$

A Simple Example of Q-Learning

Suppose: **initial state = 1**
A(1) = {go to 3, go to 5}

By random selection: **a = go to 5**
Experience = **(1, go to 5, 5,100)**

Initial Q-table

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

$$Q(1,5) = 100 + 0.8 \max(Q(5,.))$$
$$= 100 + 0.8 * 0 = 100$$

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

A Simple Example of Q-Learning

New episode: **state = 3**
A(3) = {go to 1, go to 2, go to 4}
Random selection: **a = go to 1**

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	100
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

A Simple Example of Q-Learning

New episode: **state = 3**
A(3) = {go to 1, go to 2, go to 4}
Random selection: **a = go to 1**

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	100
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Experience = **(3, go to 1, 1, 0)**
A(1) = {go to 3, go to 5}

$$Q(3,1) = 0 + 0.8 \max(Q(1,3), Q(1,5))$$
$$= 0 + 0.8 * 100 = 80$$

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	100
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

A Simple Example of Q-Learning

Now **state = 1**
A(1) = {go to 3, go to 5}
ϵ-greedy: **a = go to 5**

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	100
2	0	0	0	0	0	0
3	0	80	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

A Simple Example of Q-Learning

Now **state = 1**

A(1) = {go to 3, go to 5}

ϵ-greedy: **a = go to 5**

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	100
2	0	0	0	0	0	0
3	0	80	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Experience = **(1, go to 5, 5,100)**

A(5) = { }

$$Q(1,5) = 100 + 0.8 \max(Q(5,.))$$
$$= 100 + 0.8 * 0 = 100$$

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	80	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

A Simple Example of Q-Learning

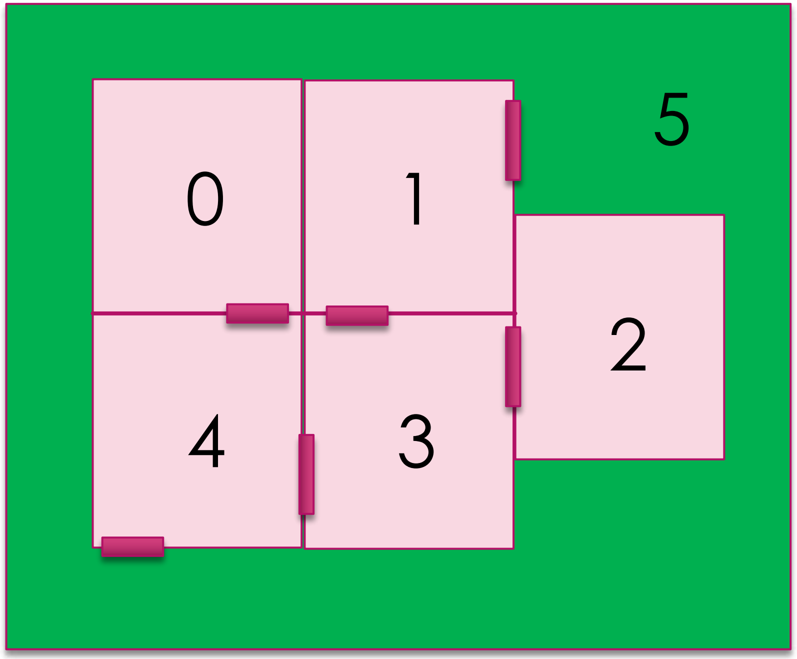
If our agent learns more through further episodes, it will finally reach convergence values in Q-table like:

s \ a	0	1	2	3	4	5
0	0	0	0	0	80	0
1	0	0	0	64	0	100
2	0	0	0	64	0	0
3	0	80	51	0	80	0
4	64	0	0	64	0	100
5	0	0	0	0	0	0

$V^*(s)$ = max_a $Q(s,a)$
$\pi^*(0)$ = arg max_a $Q(s,a)$

$V^*(0)=80$ $\pi^*(0)$ = go to 4
$V^*(1)=100$ $\pi^*(1)$ = go to 5
$V^*(2)=64$ $\pi^*(2)$ = go to 3
$V^*(3)=80$ $\pi^*(3)$ = go to 4 or 1
$V^*(4)=100$ $\pi^*(4)$ = go to 5
$V^*(5)=0$ $\pi^*(5)$ = { } (goal)

A Simple Example of Q-Learning



$\equiv$