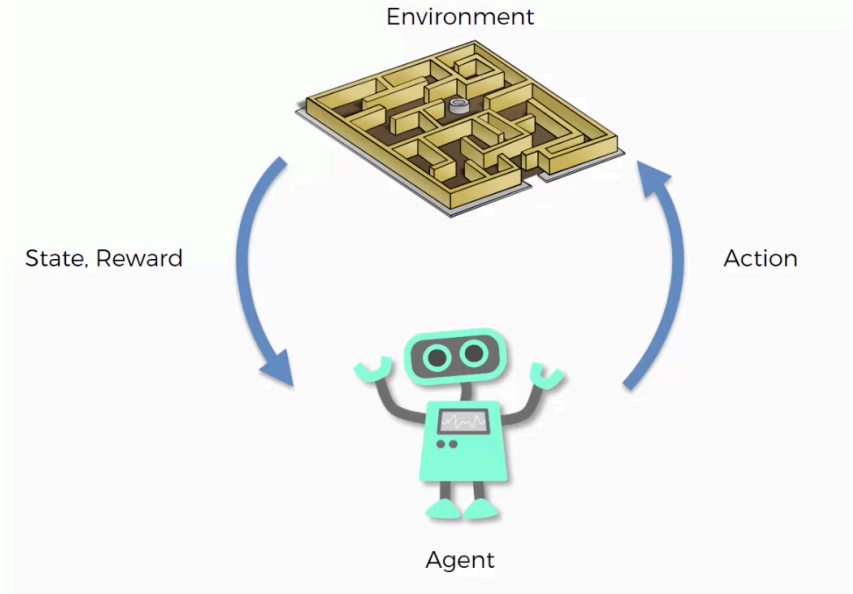
RL Plan of Attack

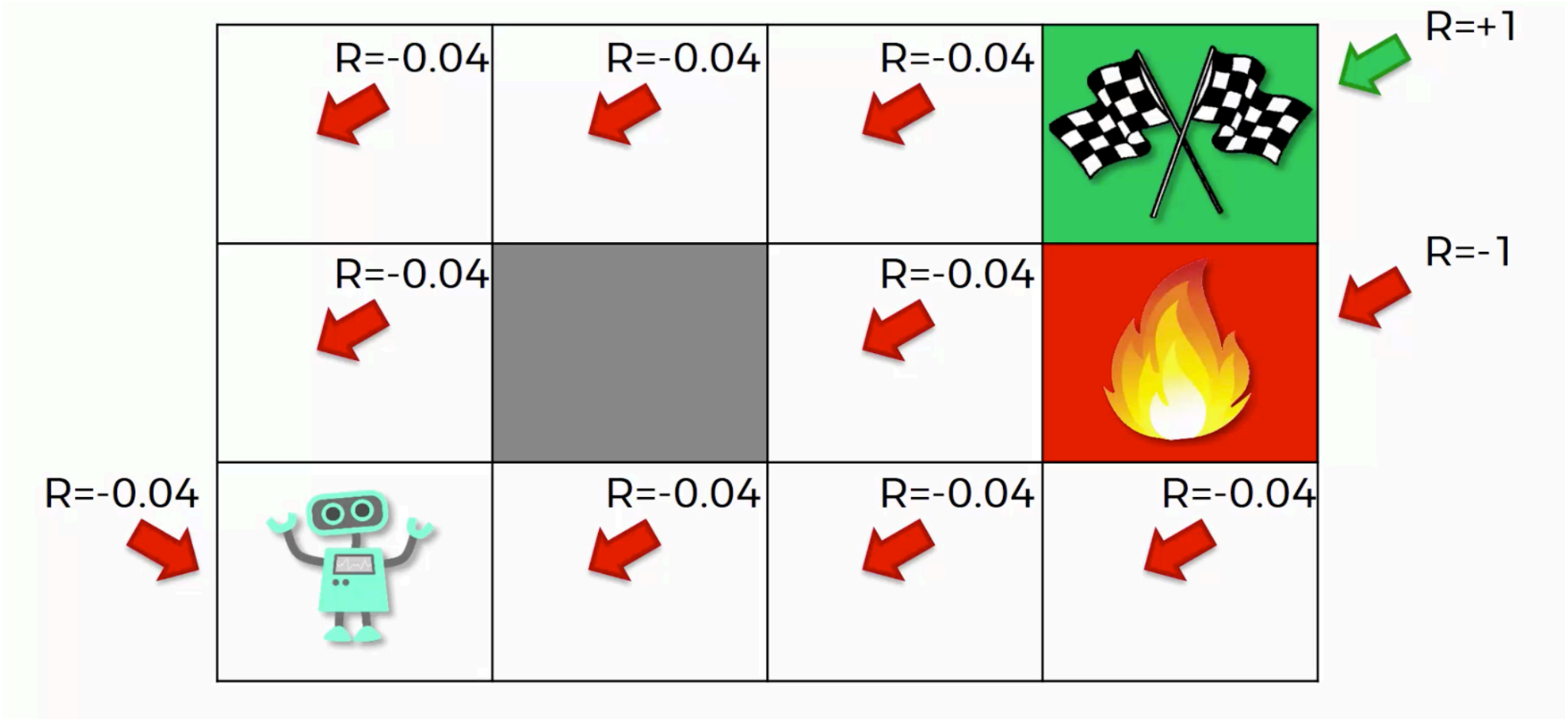
1. What is Reinforcement Learning? (RL)
2. The Bellman Equation
3. The ”Plan”
4. Markov Decision Process (MDP)
5. “Policy” vs. “Plan”
6. Living Penalty
7. Q-Learning Intuition
8. Temporal Difference
9. Deep Q-Learning (Learning & Acting)

6. Living Penalty

$$V(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \right)$$

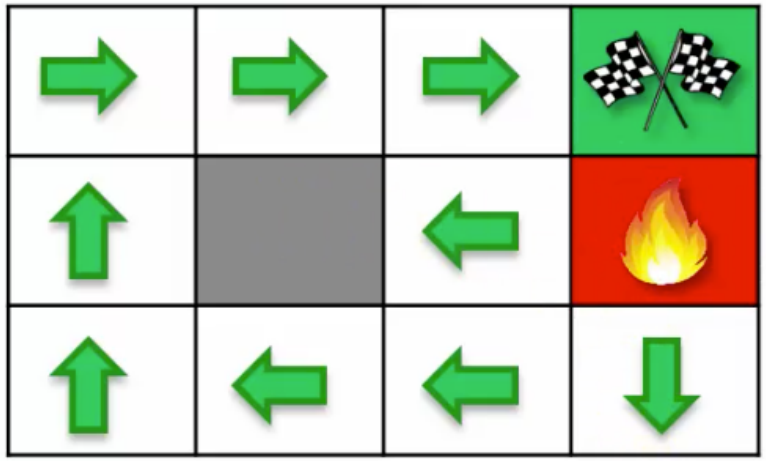


6. Living Penalty

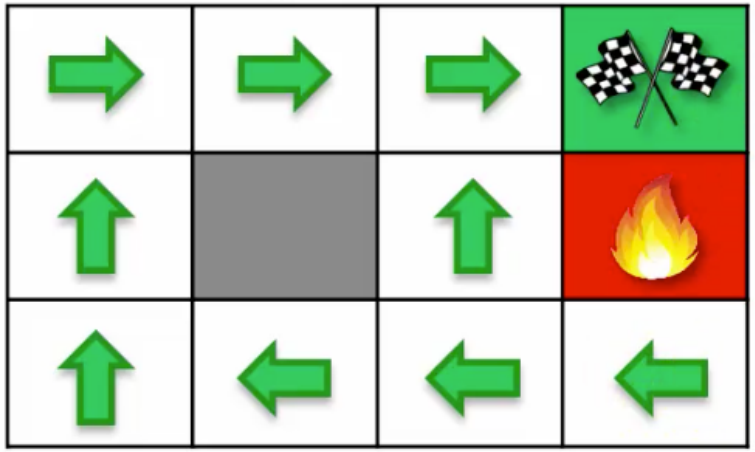


6. Living Penalty

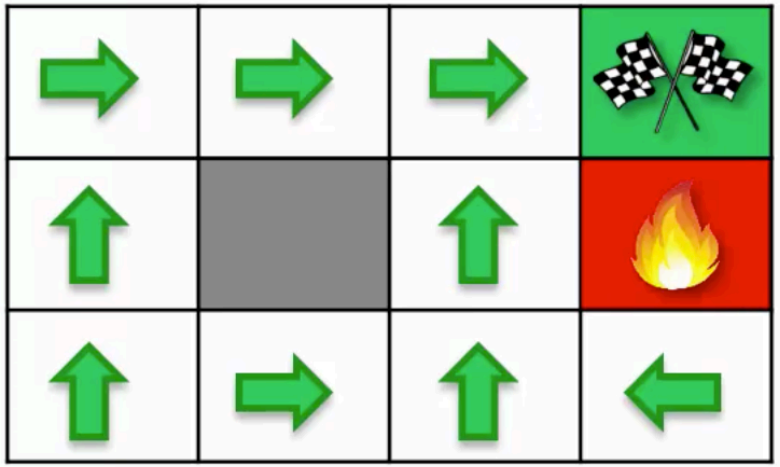
R(s)=0



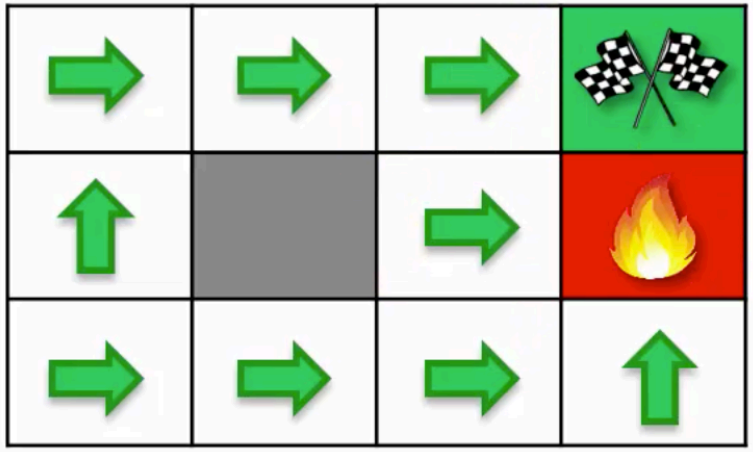
R(s)=-0.04



R(s)=-0.5



R(s)=-2.0



RL Plan of Attack

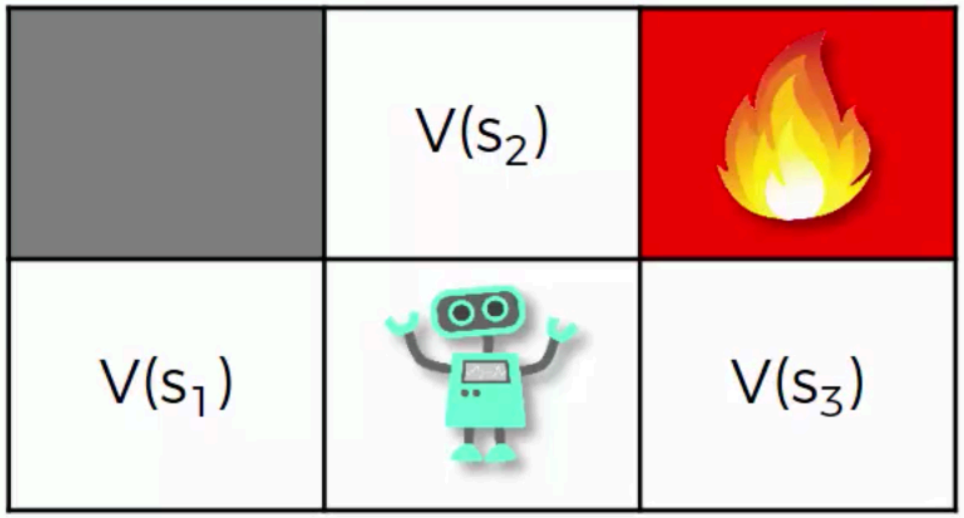
1. What is Reinforcement Learning? (RL)
2. The Bellman Equation
3. The ”Plan”
4. Markov Decision Process (MDP)
5. “Policy” vs. “Plan”
6. Living Penalty
7. Q-Learning Intuition
8. Temporal Difference
9. Deep Q-Learning (Learning & Acting)

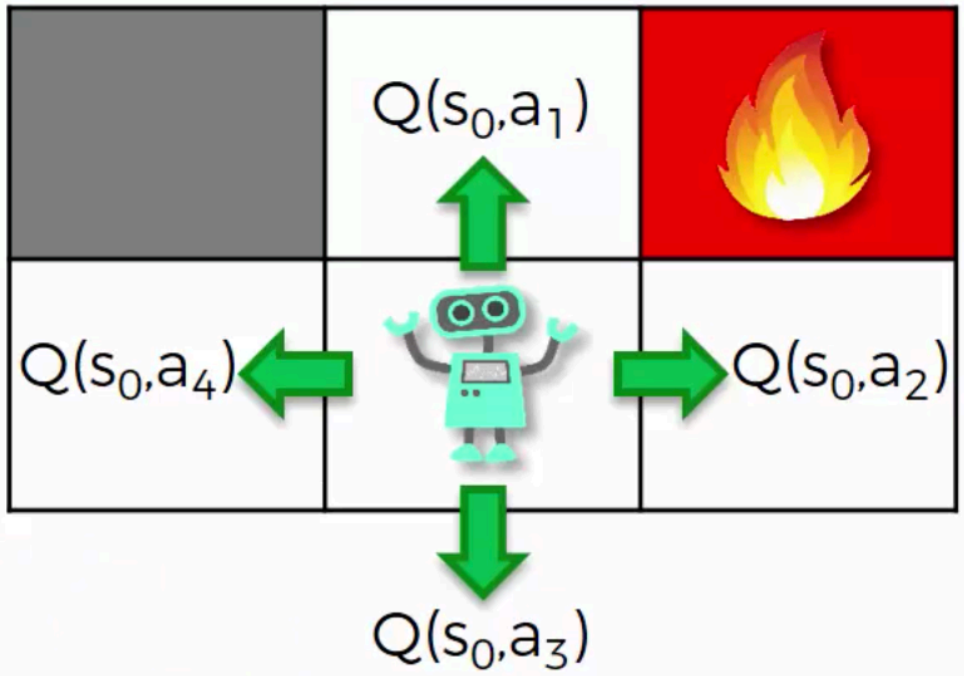
7. Q-Learning Intuition

$$V(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \right)$$

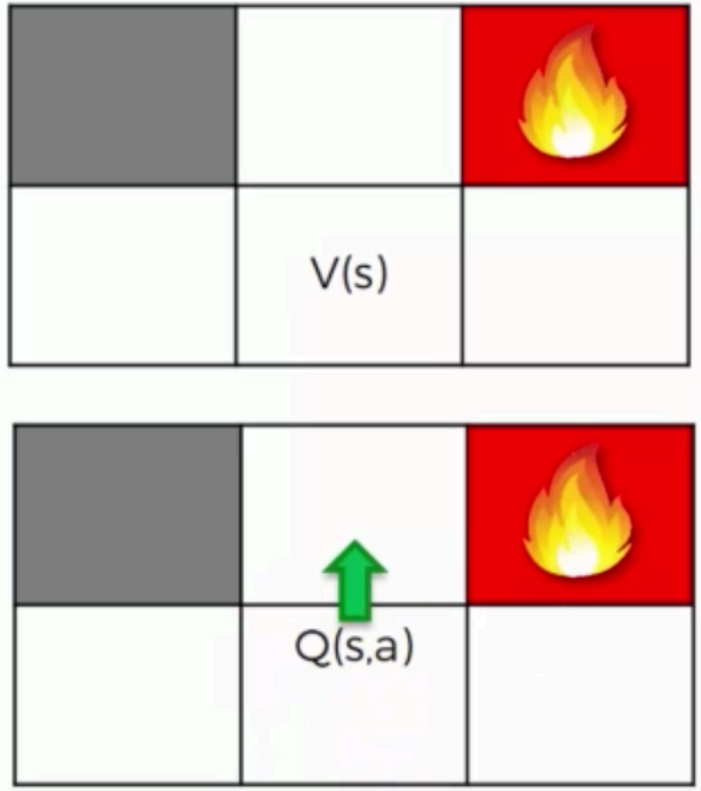
WHERE'S THE Q?

7. Q-Learning Intuition





7. Q-Learning Intuition



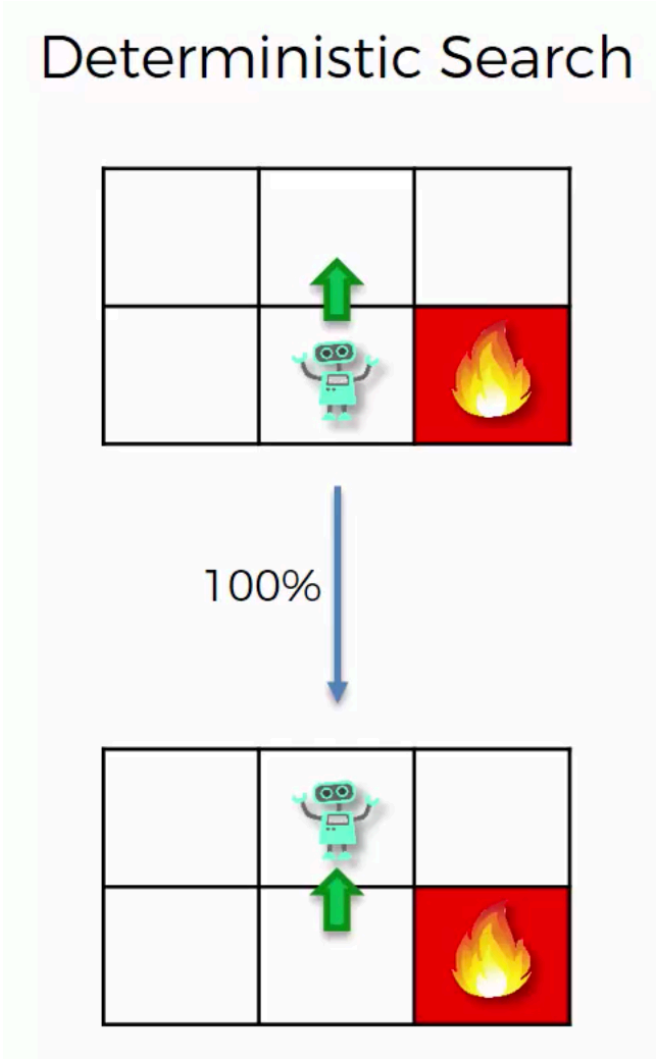
$$V(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \right)$$

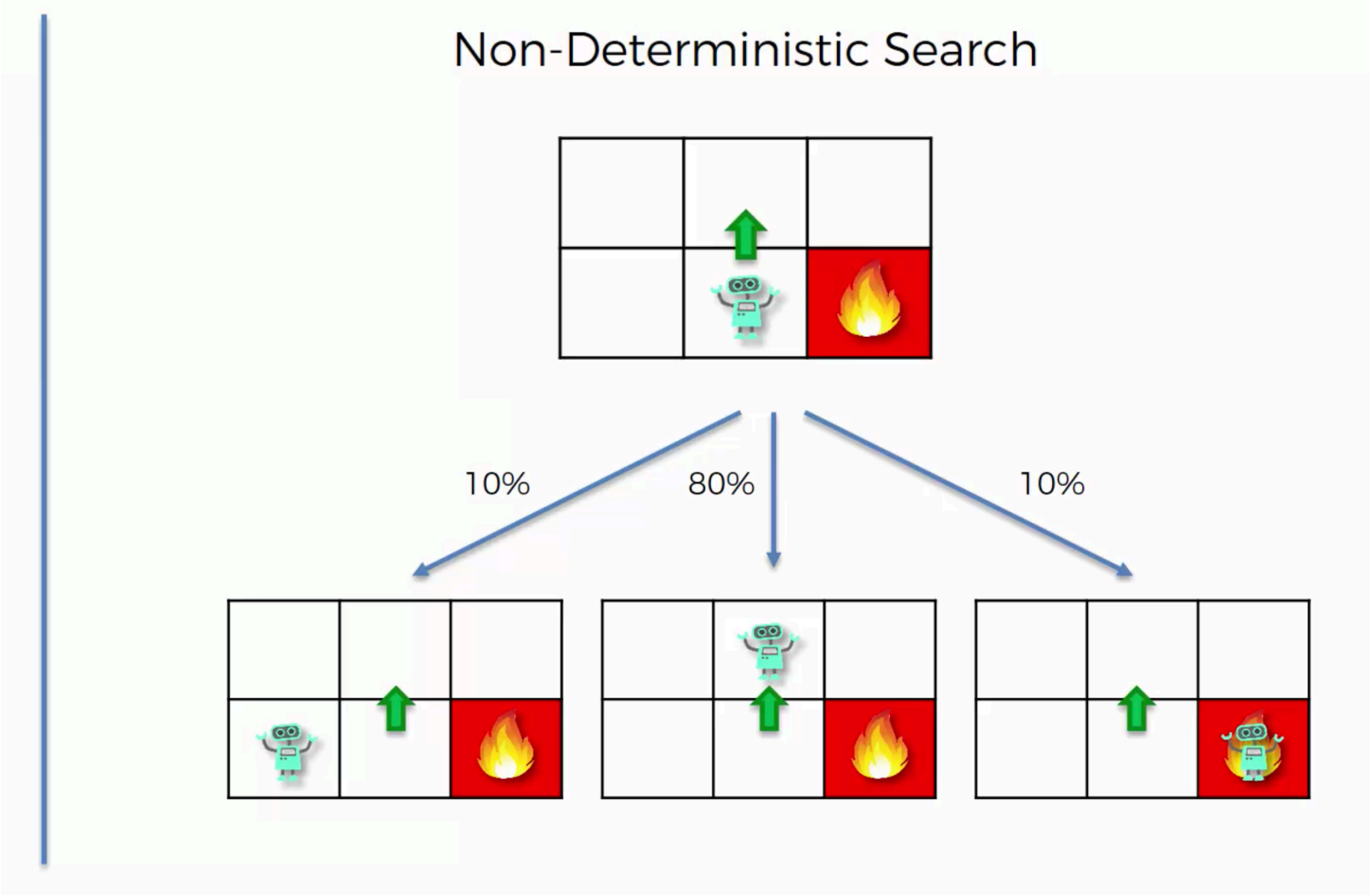
$$Q(s,a) = R(s,a) + \gamma \sum_{s'} \left(P(s,a,s') \max_{a'} Q(s',a') \right)$$

RL Plan of Attack

1. What is Reinforcement Learning? (RL)
2. The Bellman Equation
3. The ”Plan”
4. Markov Decision Process (MDP)
5. “Policy” vs. “Plan”
6. Living Penalty
7. Q-Learning Intuition
8. Temporal Difference
9. Deep Q-Learning (Learning & Acting)

8. Temporal Difference





8. Temporal Difference

V=0.81	V=0.9	V=1	
V=0.73		V=0.9	
V=0.66	V=0.73	V=0.81	V=0.73

8. Temporal Difference

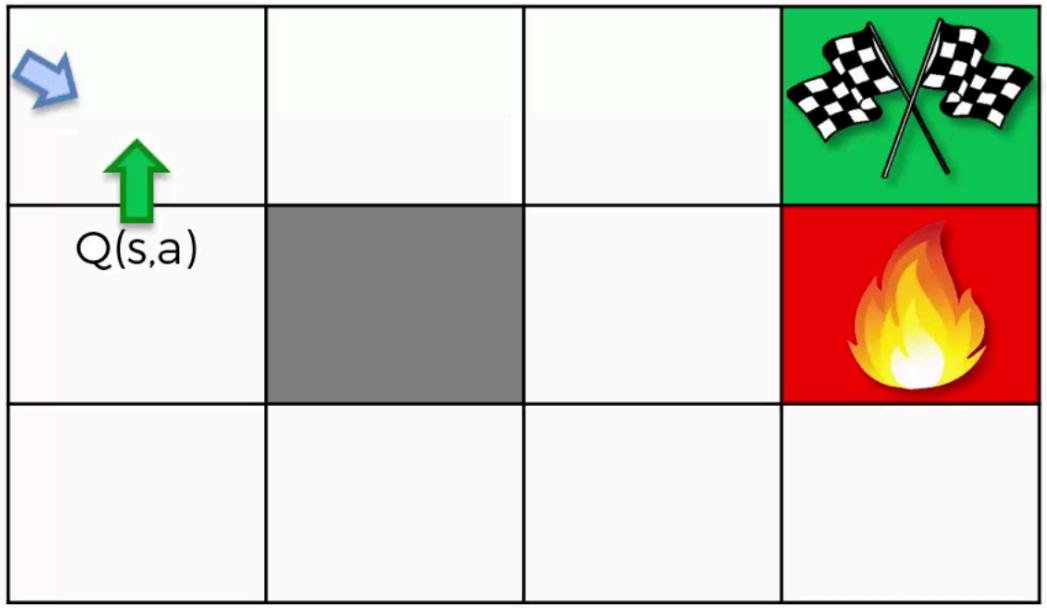
V=0.71	V=0.74	V=0.86	
V=0.63		V=0.39	
V=0.55	V=0.46	V=0.36	V=0.22

8. Temporal Difference

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} \left(P(s,a,s') \max_{a'} Q(s',a') \right)$$

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q(s',a')$$

8. Temporal Difference



Before:

$Q(s,a)$

After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

$$TD(a,s) = \overbrace{R(s,a) + \gamma \max_{a'} Q(s',a')}^{} - \overbrace{Q(s,a)}^{}$$

8. Temporal Difference

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q_{t-1}(s,a)$$

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha TD_t(a,s)$$

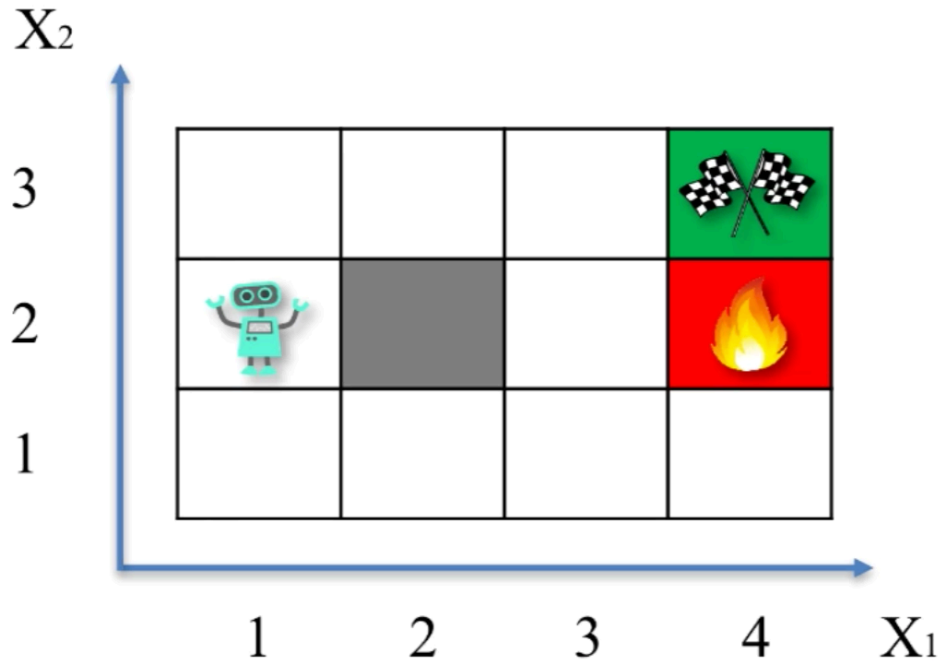
$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q(s',a') - Q_{t-1}(s,a) \right)$$

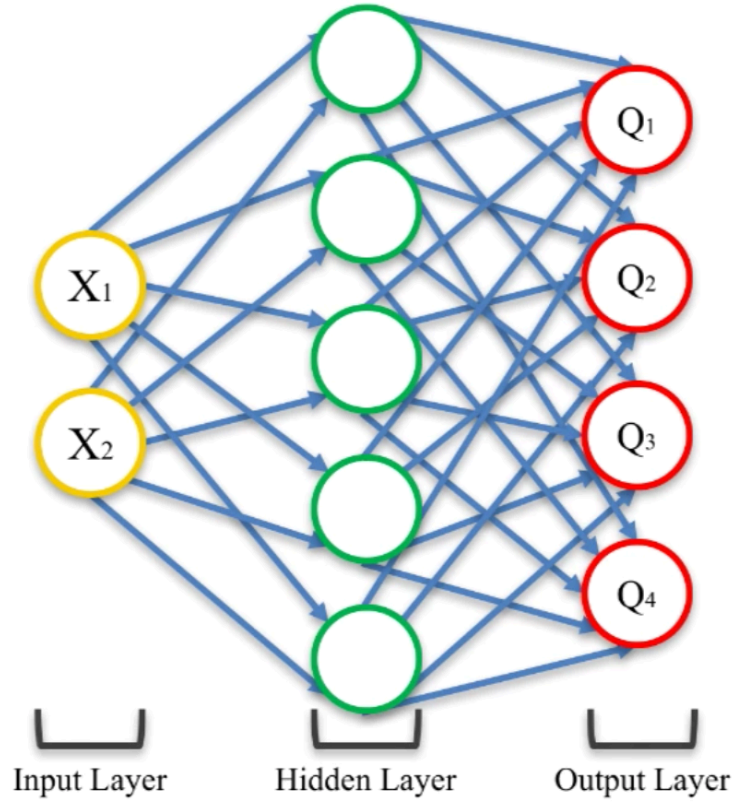
RL Plan of Attack

1. What is Reinforcement Learning? (RL)
2. The Bellman Equation
3. The ”Plan”
4. Markov Decision Process (MDP)
5. “Policy” vs. “Plan”
6. Living Penalty
7. Q-Learning Intuition
8. Temporal Difference
9. Deep Q-Learning (Learning & Acting)

Recap

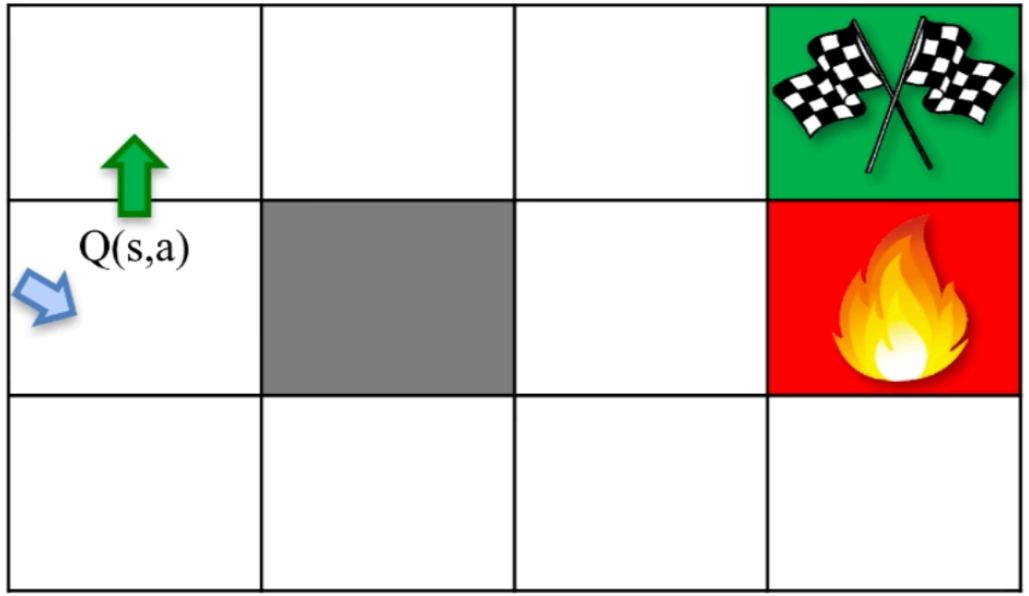
9. Deep Q-Learning (Learning & Acting)





9. Deep Q-Learning (Learning & Acting)

The Essene of simple Q-Learning



Before:

$Q(s,a)$

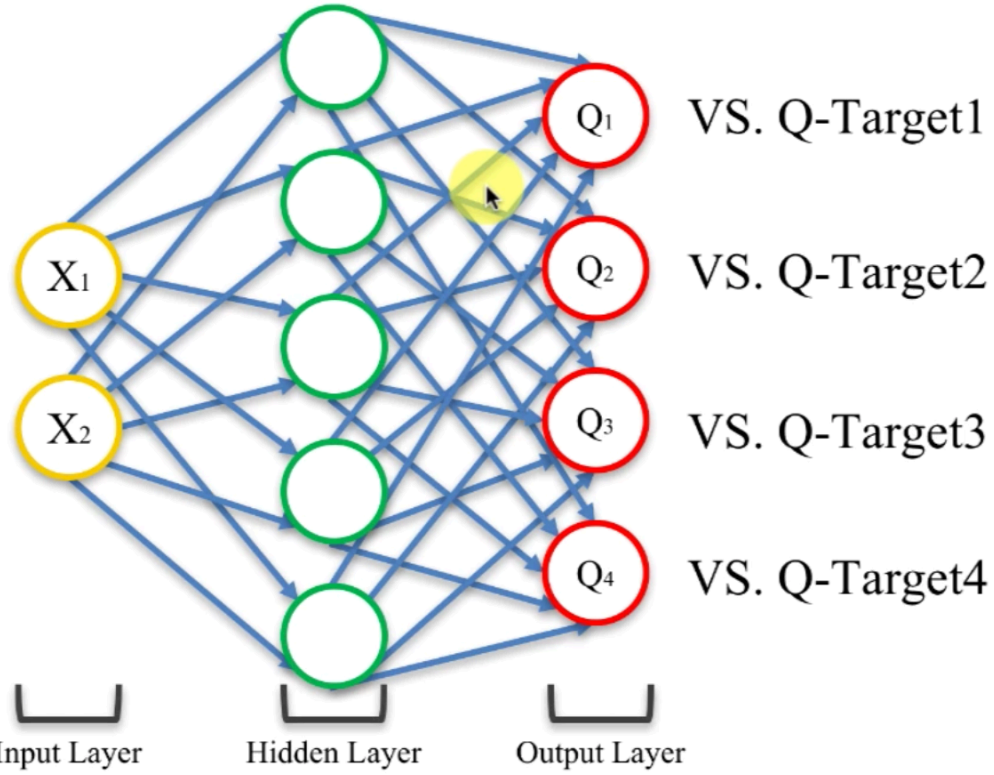
After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

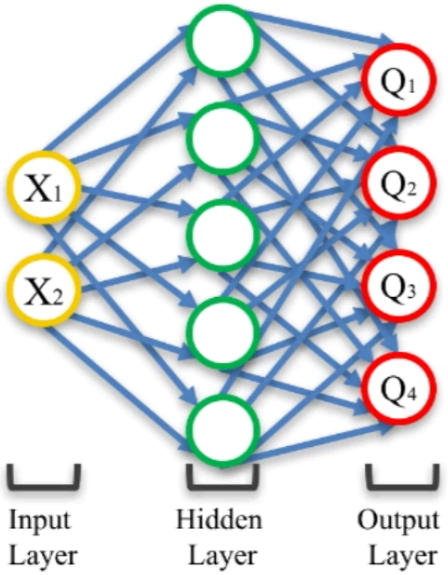
9. Deep Q-Learning (Learning & Acting)

Learning:



$$L = \sum (Q\text{–}Target\ - Q)^2$$

9. Deep Q-Learning (Learning & Acting)



Learning:

VS. Q-Target1

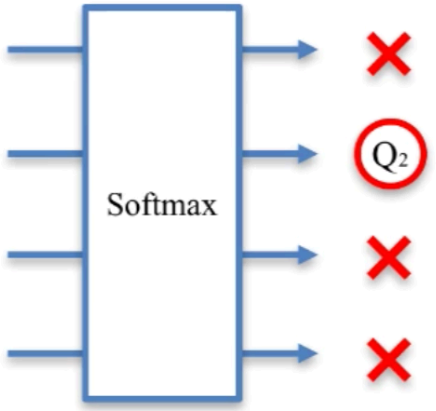
VS. Q-Target2

VS. Q-Target3

VS. Q-Target4

$$L = \sum (Q\text{–}Target - Q)^2$$

Acting:



Weights updated Every time the state is changed

Further studies…

- Experience Replay
- Deep Convolutional Q-Learning
- A3C
- And much more topics to study…

Code Examples:

- https://keras.io/examples/rl/deep_q_network_breakout/
- https://towardsdatascience.com/reinforcement-learning-w-keras-openai-dqns-1eed3a5338c