



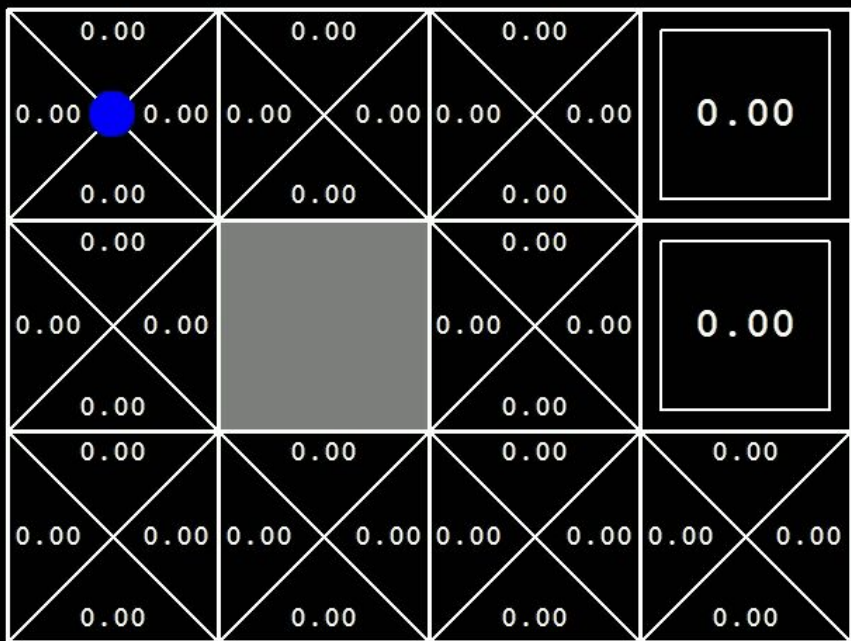
# Q-Learning

Como aprender quais ações tomar



**GRUPO**  
**TURING**

# Objetivo



CURRENT Q-VALUES

Queremos estimar o valor  $q$  de cada ação, para poder escolher as melhores.

Para isto, precisamos explorar as ações do ambiente em cada estado para descobrir suas recompensas médias.

# Dilema

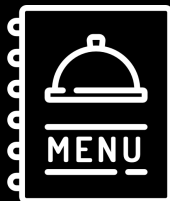
---

## Exploração

Aproveitar ações conhecidas

Maior recompensa com o conhecimento que já possui

Não garante que se trata das melhores escolhas



## Exploração

Tomar novas ações para explorar estados diferentes

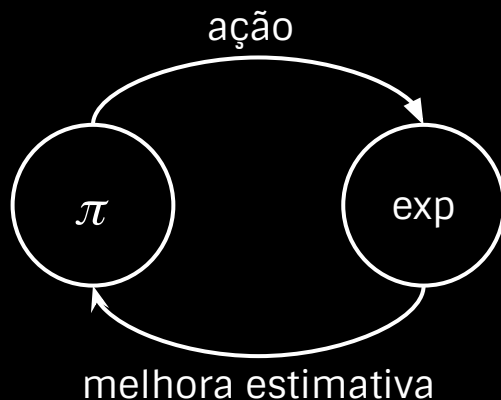
Pode levar a caminhos melhores

Menores recompensas a curto prazo

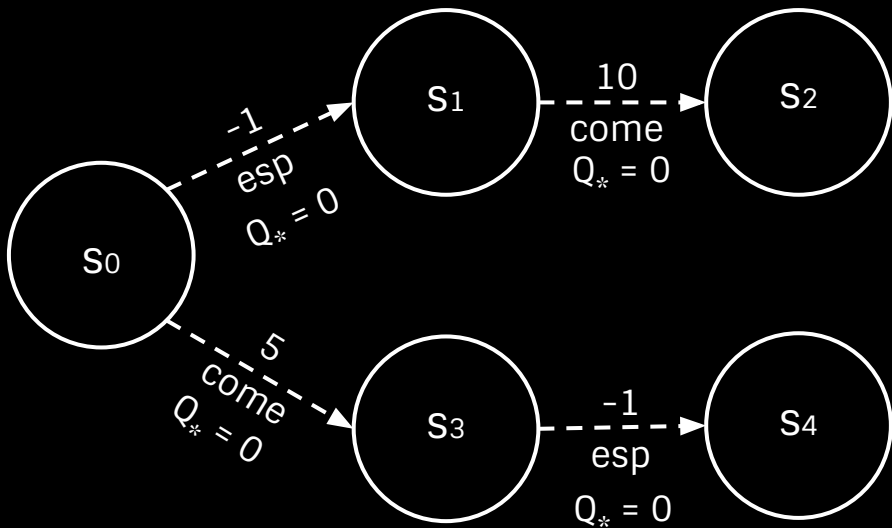
Objetivo: buscar equilíbrio entre exploração e exploração

# Q-Learning

$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$

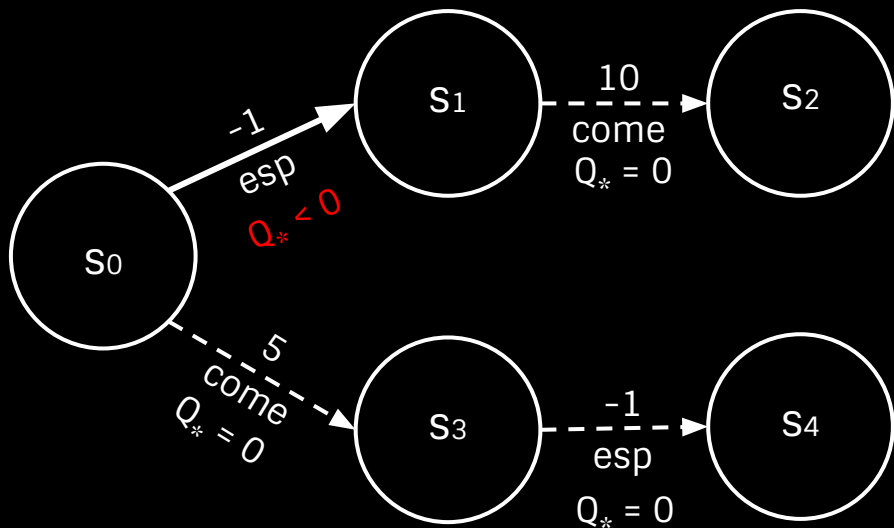


# Exemplo



$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$

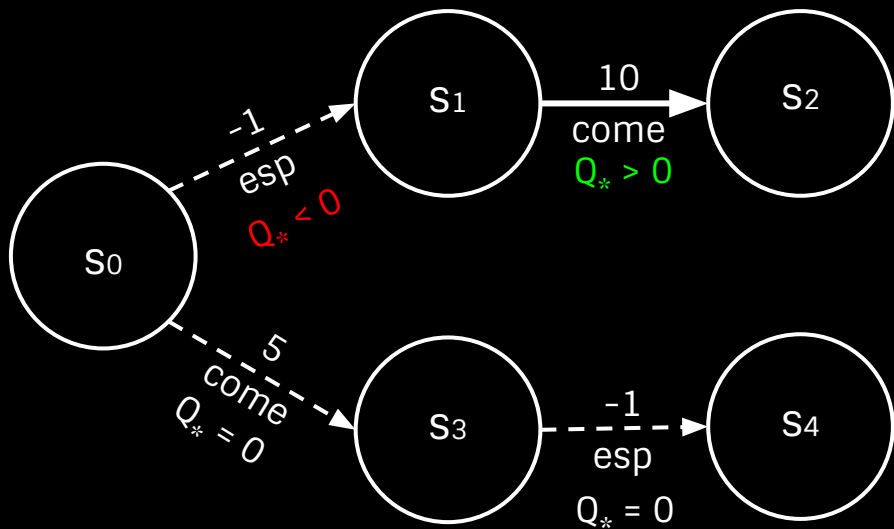
# Exemplo



$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$



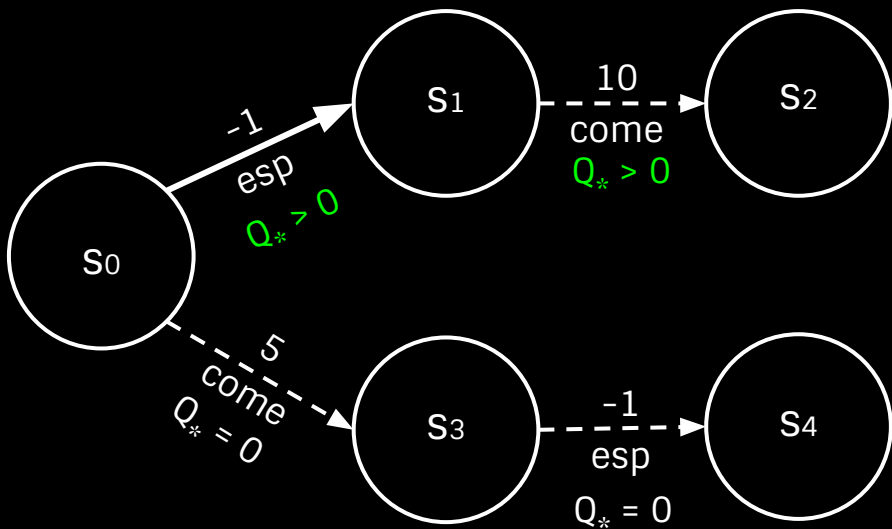
# Exemplo



$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$



# Exemplo

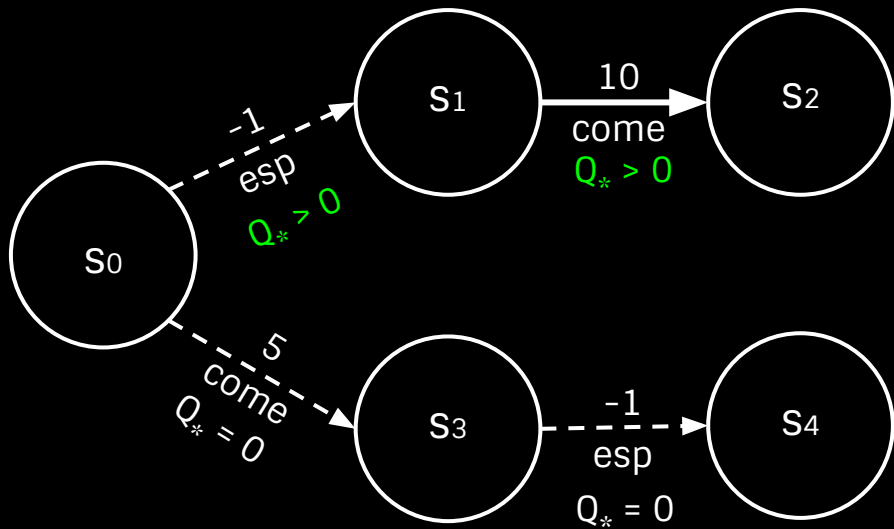


$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$





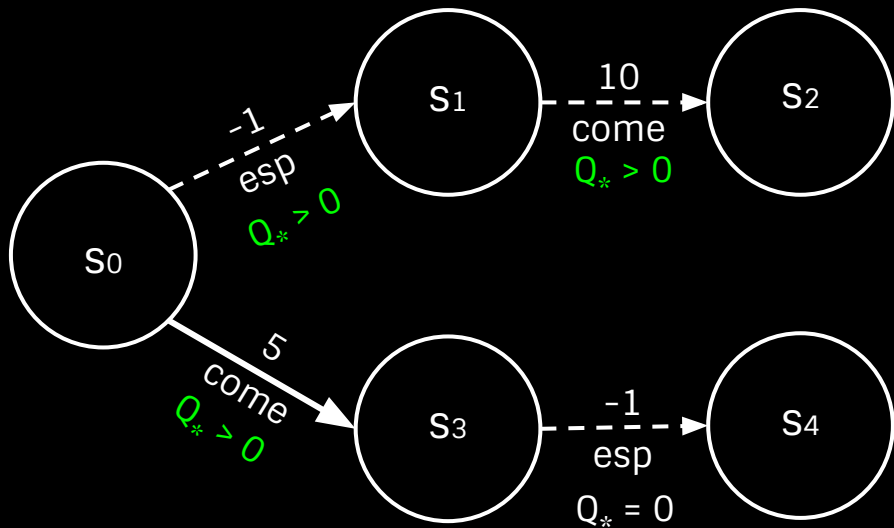
# Exemplo



$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$



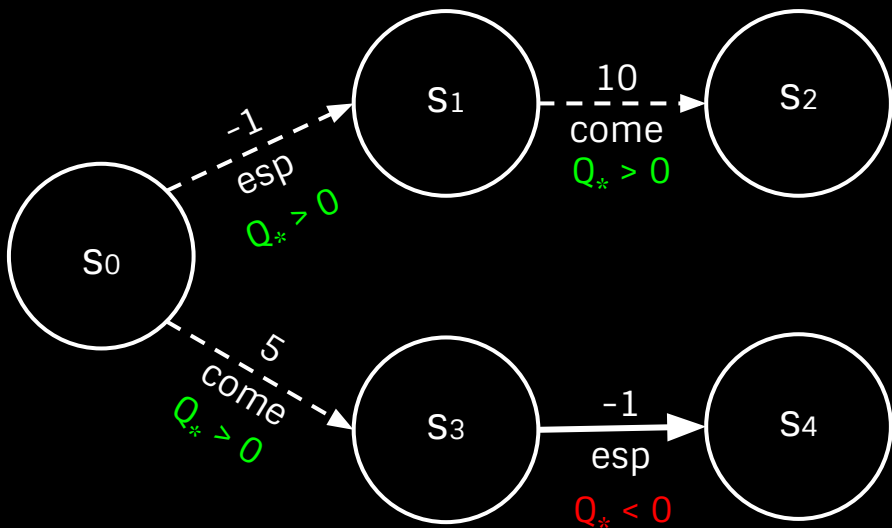
# Exemplo



$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$



# Exemplo



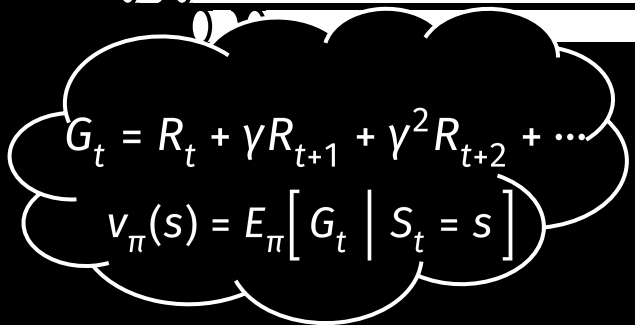
$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_*^{\text{nov}}(s, a)$$



# Como estimar os valores Q?

**GRUPO**  
**TURING**

# Equação de Bellman


$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

$$v_{\pi}(s) = E_{\pi}[G_t \mid S_t = s]$$

$$q_{*}(s, a) = E_{\pi^{*}}[G_t \mid S_t = s, A_t = a]$$

$$= E_{\pi^{*}}[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots \mid S_t = s, A_t = a]$$
 Retorno no instante t

$$= E_{\pi^{*}}[R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \dots) \mid S_t = s, A_t = a]$$
 Isolando o  $\gamma$

$$= E_{\pi^{*}}[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a]$$
 Retorno no instante t+1

$$= E_{\pi^{*}}[R_t + \gamma v_{*}(S_{t+1}) \mid S_t = s, A_t = a]$$
 Tomando o valor médio de  $G_{t+1}$

$$= E_{\pi^{*}}[R_t + \gamma \max_{a'} q_{*}(S_{t+1}, a') \mid S_t = s, A_t = a]$$
 Como o agente sempre toma a melhor ação

# Bootstrapping

Como calcular essa média?

$$q_*(s, a) = E_{\pi^*} \left[ R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a \right]$$

Uma estimativa bem grosseira:  
podemos considerar apenas uma transição

$$Q_{\text{bootstrap}}(s, a) = \underbrace{r}_{\text{recompensa}} + \underbrace{\gamma}_{\text{fator de desconto}} \cdot \underbrace{\max_{a'} Q_*(s', a')}_{\text{estimativa do valor futuro \u00f3timo}}$$

# Aplicação das conclusões: Q-learning

**GRUPO**  
**TURING**

# Q-Learning

Bootstrap gera estimativas ruins,  
que variam muito

$$Q_{\text{bootstrap}}(s, a) = \underbrace{r}_{\text{recompensa}} + \underbrace{\gamma}_{\text{fator de desconto}} \cdot \underbrace{\max_{a'} Q_*(s', a')}_{\text{estimativa do valor futuro \u00f3timo}}$$

Podemos “estabilizar” esses valores usando uma m\u00e9dia ponderada:

$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_{\text{bootstrap}}(s, a)$$

$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot \underbrace{Q_*(s, a)}_{\text{valor antigo}} + \underbrace{\alpha}_{\text{taxa de aprendizado}} \cdot \underbrace{\left( r + \gamma \max_{a'} Q_*(s', a') \right)}_{Q_{\text{bootstrap}}(s, a)}$$



# Q-Learning

$$Q_*(s, a) \leftarrow (1 - \alpha) \cdot Q_*(s, a) + \alpha \cdot Q_{\text{bootstrap}}(s, a)$$

$$Q_*(s, a) \leftarrow \underbrace{(1 - \alpha) \cdot Q_*(s, a)}_{\text{valor antigo}} + \underbrace{\alpha}_{\text{taxa de aprendizado}} \cdot \underbrace{\left( r + \gamma \max_{a'} Q_*(s', a') \right)}_{Q_{\text{bootstrap}}(s, a)}$$

média ponderada

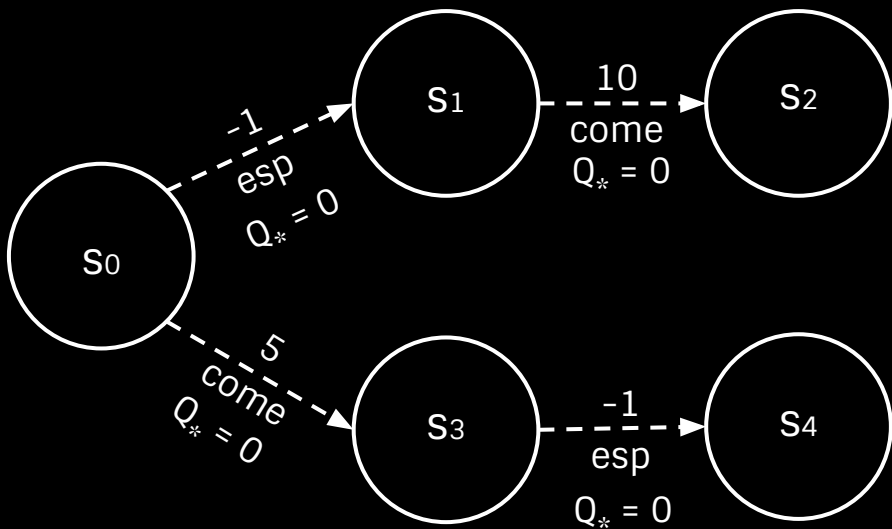
$$Q_*(s, a) \leftarrow \underbrace{Q_*(s, a)}_{\text{valor antigo}} + \underbrace{\alpha}_{\text{taxa de aprendizado}} \cdot \underbrace{\left( \underbrace{r + \gamma \max_{a'} Q_*(s', a')}_{Q_{\text{bootstrap}}(s, a)} - Q_*(s, a) \right)}_{\text{erro}}$$

minimização do erro

# Exemplo

$$\alpha = 0.5$$

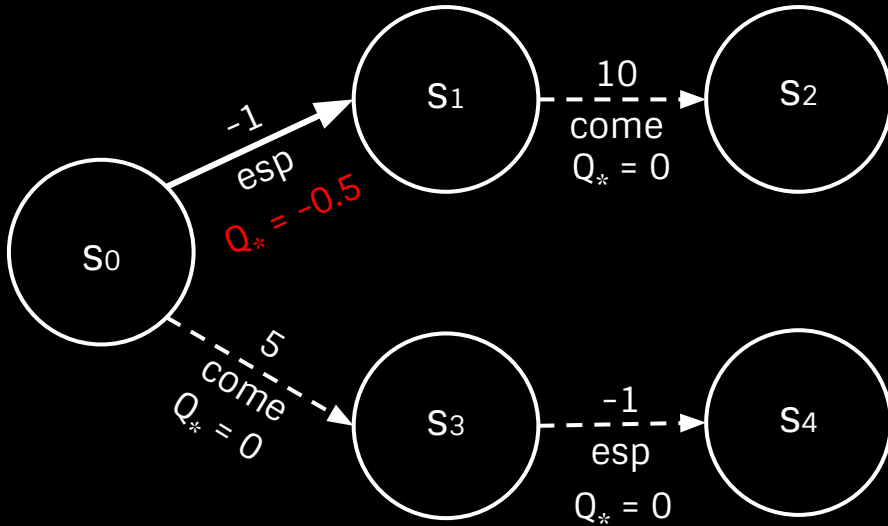
$$\gamma = 0.9$$



# Exemplo (ep. 1)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{boot}(s_0, esp) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 0 = -1$$

$$Q_*(s_0, esp) \leftarrow (1 - \alpha) \cdot Q_*(s_0, esp) + \alpha \cdot Q_{boot}(s_0, esp)$$

$$Q_*(s_0, esp) \leftarrow 0.5 \cdot 0 + 0.5 \cdot (-1)$$

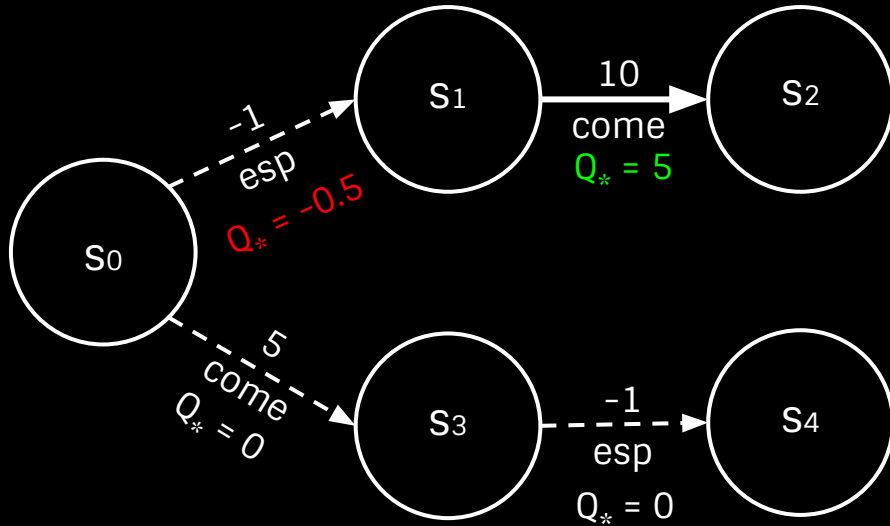
$$Q_*(s_0, esp) \leftarrow -0.5$$



# Exemplo (ep. 1)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{\text{boot}}(s_0, \text{esp}) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 0 = -1$$

$$Q_*(s_0, \text{esp}) \leftarrow (1 - \alpha) \cdot Q_*(s_0, \text{esp}) + \alpha \cdot Q_{\text{boot}}(s_0, \text{esp})$$

$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot 0 + 0.5 \cdot (-1)$$

$$Q_*(s_0, \text{esp}) \leftarrow -0.5$$

$$Q_{\text{boot}}(s_1, \text{come}) = r + \gamma \max_{a'} Q_*(s_2, a') = 10 + 0.9 \cdot 0 = 10$$

$$Q_*(s_1, \text{come}) \leftarrow (1 - \alpha) \cdot Q_*(s_1, \text{come}) + \alpha \cdot Q_{\text{boot}}(s_1, \text{come})$$

$$Q_*(s_1, \text{come}) \leftarrow 0.5 \cdot 0 + 0.5 \cdot 10$$

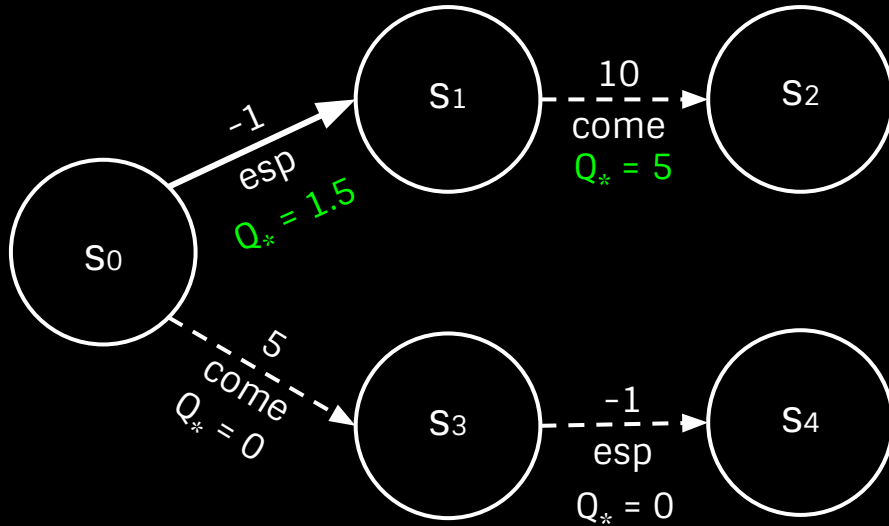
$$Q_*(s_1, \text{come}) \leftarrow 5$$



# Exemplo (ep. 2)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{\text{boot}}(s_0, \text{esp}) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 5 = 3.5$$

$$Q_*(s_0, \text{esp}) \leftarrow (1 - \alpha) \cdot Q_*(s_0, \text{esp}) + \alpha \cdot Q_{\text{boot}}(s_0, \text{esp})$$

$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot (-0.5) + 0.5 \cdot 3.5$$

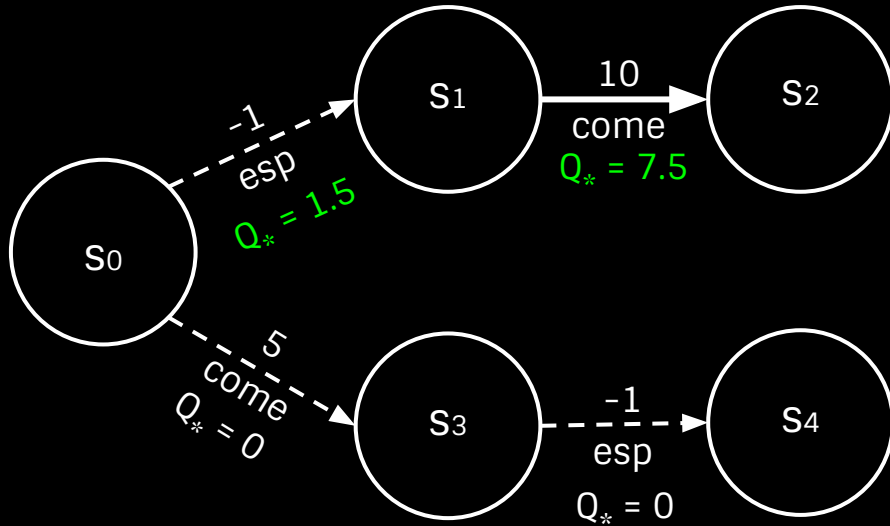
$$Q_*(s_0, \text{esp}) \leftarrow 1.5$$



# Exemplo (ep. 2)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{\text{boot}}(s_0, \text{esp}) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 5 = 3.5$$

$$Q_*(s_0, \text{esp}) \leftarrow (1 - \alpha) \cdot Q_*(s_0, \text{esp}) + \alpha \cdot Q_{\text{boot}}(s_0, \text{esp})$$

$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot (-0.5) + 0.5 \cdot 3.5$$

$$Q_*(s_0, \text{esp}) \leftarrow 1.5$$

$$Q_{\text{boot}}(s_1, \text{come}) = r + \gamma \max_{a'} Q_*(s_2, a') = 10 + 0.9 \cdot 0 = 10$$

$$Q_*(s_1, \text{come}) \leftarrow (1 - \alpha) \cdot Q_*(s_1, \text{come}) + \alpha \cdot Q_{\text{boot}}(s_1, \text{come})$$

$$Q_*(s_1, \text{come}) \leftarrow 0.5 \cdot 5 + 0.5 \cdot 10$$

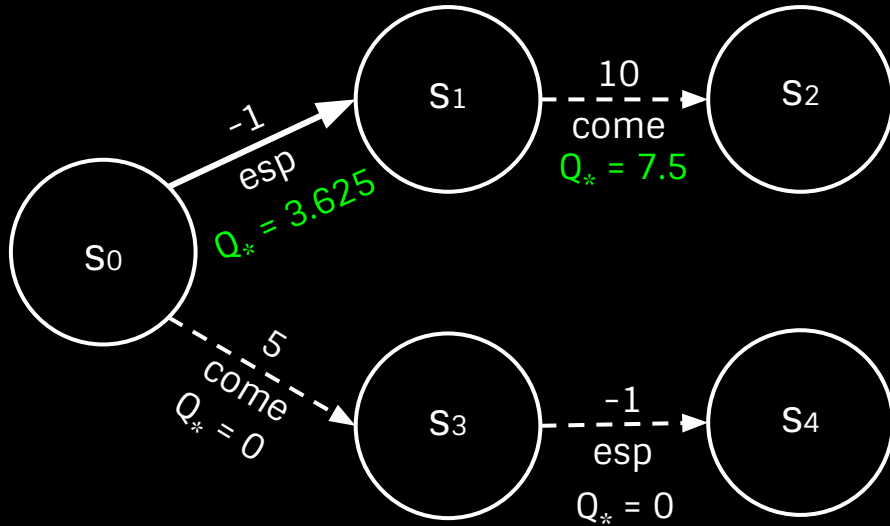
$$Q_*(s_1, \text{come}) \leftarrow 7.5$$



# Exemplo (ep. 3)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{\text{boot}}(s_0, \text{esp}) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 7.5 = 5.75$$

$$Q_*(s_0, \text{esp}) \leftarrow (1 - \alpha) \cdot Q_*(s_0, \text{esp}) + \alpha \cdot Q_{\text{boot}}(s_0, \text{esp})$$

$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot 1.5 + 0.5 \cdot 5.75$$

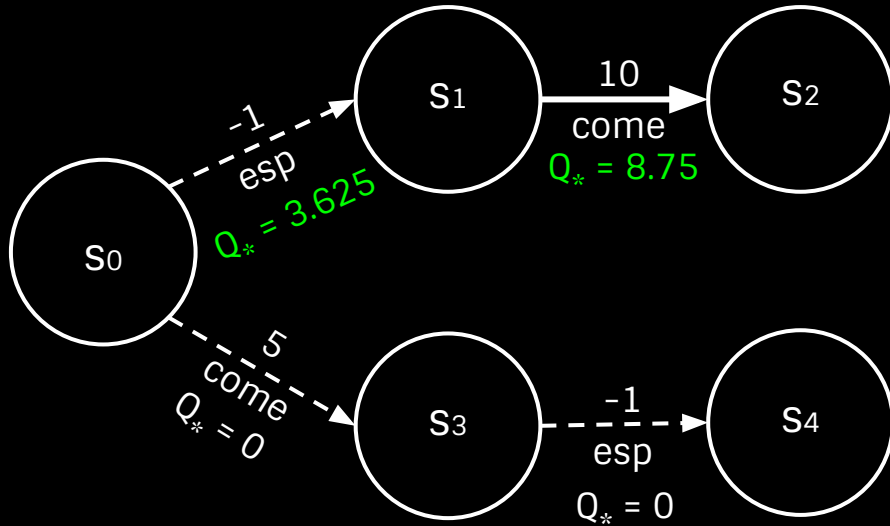
$$Q_*(s_0, \text{esp}) \leftarrow 3.625$$



# Exemplo (ep. 3)

$$\alpha = 0.5$$

$$\gamma = 0.9$$



$$Q_{\text{boot}}(s_0, \text{esp}) = r + \gamma \max_{a'} Q_*(s_1, a') = -1 + 0.9 \cdot 7.5 = 5.75$$

$$Q_*(s_0, \text{esp}) \leftarrow (1 - \alpha) \cdot Q_*(s_0, \text{esp}) + \alpha \cdot Q_{\text{boot}}(s_0, \text{esp})$$

$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot 1.5 + 0.5 \cdot 5.75$$

$$Q_*(s_0, \text{esp}) \leftarrow 3.625$$

$$Q_{\text{boot}}(s_1, \text{come}) = r + \gamma \max_{a'} Q_*(s_2, a') = 10 + 0.9 \cdot 0 = 10$$

$$Q_*(s_1, \text{come}) \leftarrow (1 - \alpha) \cdot Q_*(s_1, \text{come}) + \alpha \cdot Q_{\text{boot}}(s_1, \text{come})$$

$$Q_*(s_1, \text{come}) \leftarrow 0.5 \cdot 7.5 + 0.5 \cdot 10$$

$$Q_*(s_1, \text{come}) \leftarrow 8.75$$





# Q-Learning Tabular

Q-learning funciona armazenando as estimativas dos Q-valores numa tabela.

As estimativas da tabela são atualizadas através da já mencionada equação de Bellman.

Q	a0	a1	a2
s0	1	15	2
s1	24	5	16
s2	10	62	-7
s3	10	15	35

# Algoritmo de Q-Learning



Parâmetros: parâmetros  $\alpha, \gamma \in (0, 1]$ ,  $\varepsilon$  pequeno  $> 0$ .

Inicialize  $Q(s, a)$ , arbitrariamente, para todo  $s, a$ , exceto quando  $Q(\text{terminal}, \cdot) = 0$

Loop para cada episódio:

    Inicialize  $S$

    Loop para cada instante do episódio:

        Escolha  $A$  usando uma política derivada de  $Q$  (e.g,  $\varepsilon$ -gulosa)

        Tome a ação  $A$ , observe  $R, S'$

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_{a'} Q(S', a') - Q(S, A)]$

$S \leftarrow S'$

    até que  $S$  seja o estado terminal



# Recursos

Turing Talks: textos desde o básico de RL até algoritmos mais avançados

## Curso:

- Reinforcement Learning (Coursera)

## Repositório de RL do Grupo

Turing: explicações e implementações de diversos algoritmos

## Livro:

- Reinforcement Learning: An Introduction, 2a ed. (Sutton, Barto)



<https://linktr.ee/arielguerreiro>