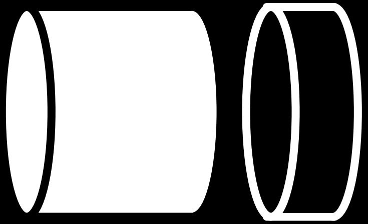
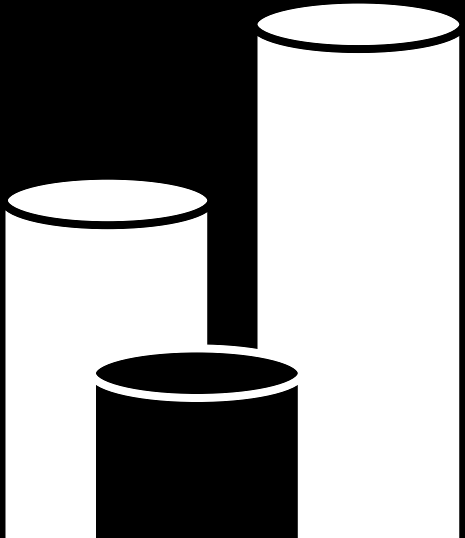
Q-Learning

Como aprender quais ações tomar

Grupo Turing

Objetivo





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

Tomar novas ações para explorar estados diferentes

Pode levar a caminhos melhores

Menores recompensas a curto prazo

Explotação

Aproveitar ações conhecidas

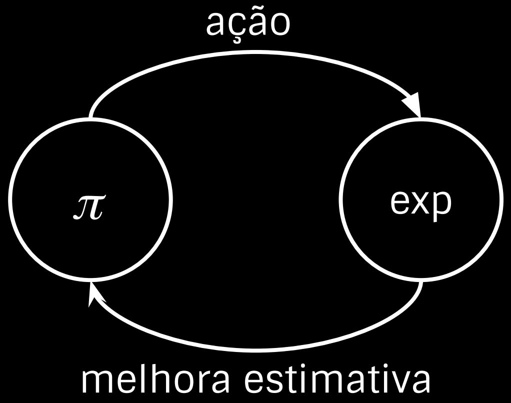
Maior recompensa com o conhecimento que já possui

Não garante que se trata das melhores escolhas

Objetivo: buscar equilíbrio entre exploração e explotação

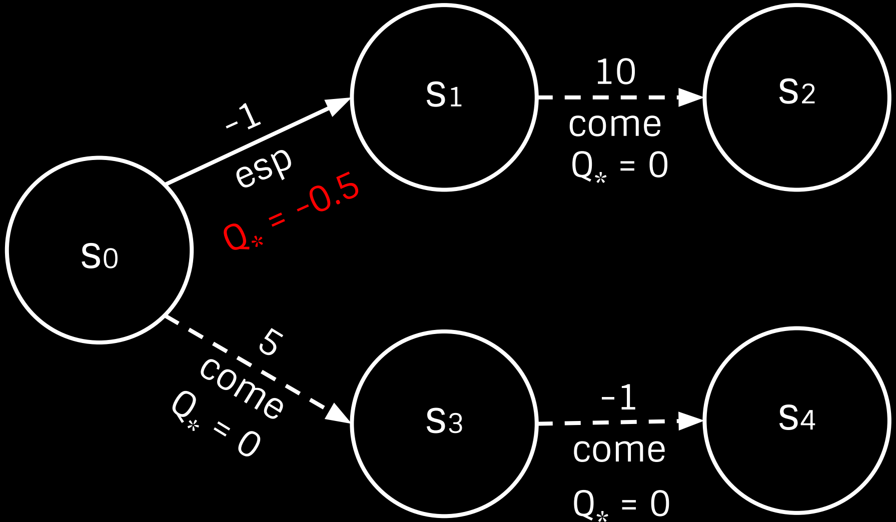
Q-Learning

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

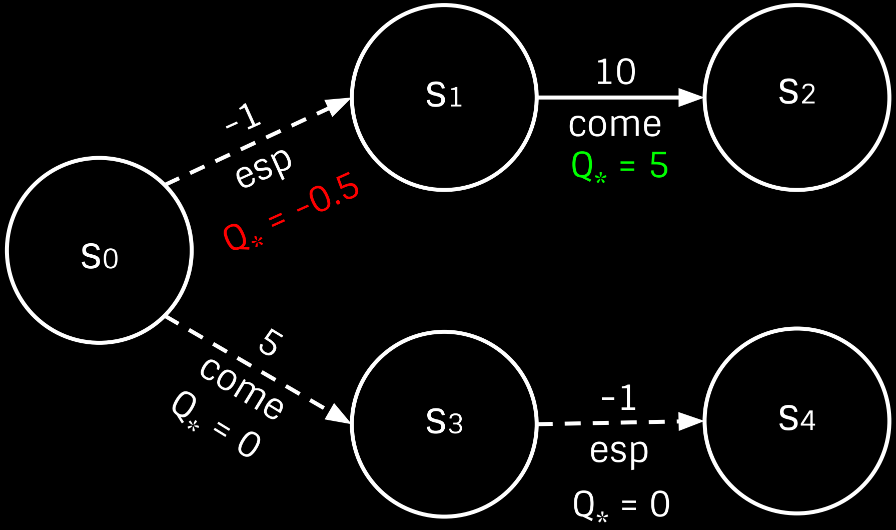


Exemplo

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

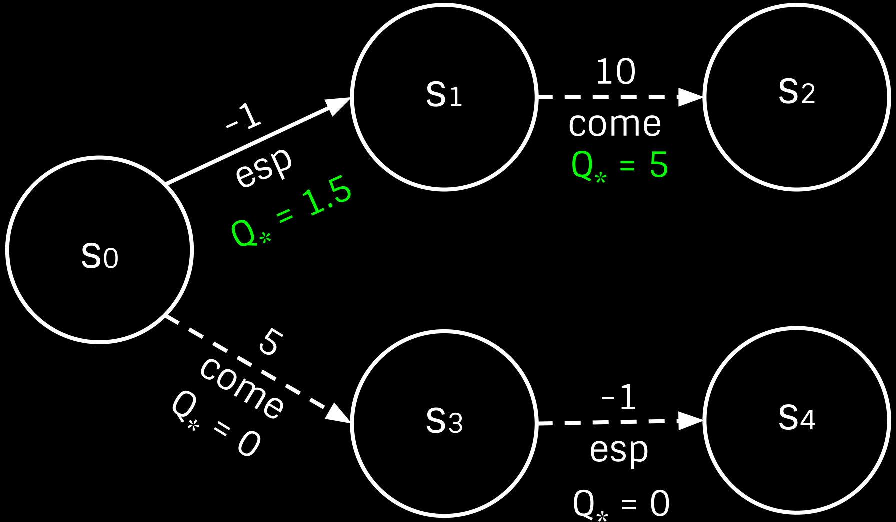


Exemplo



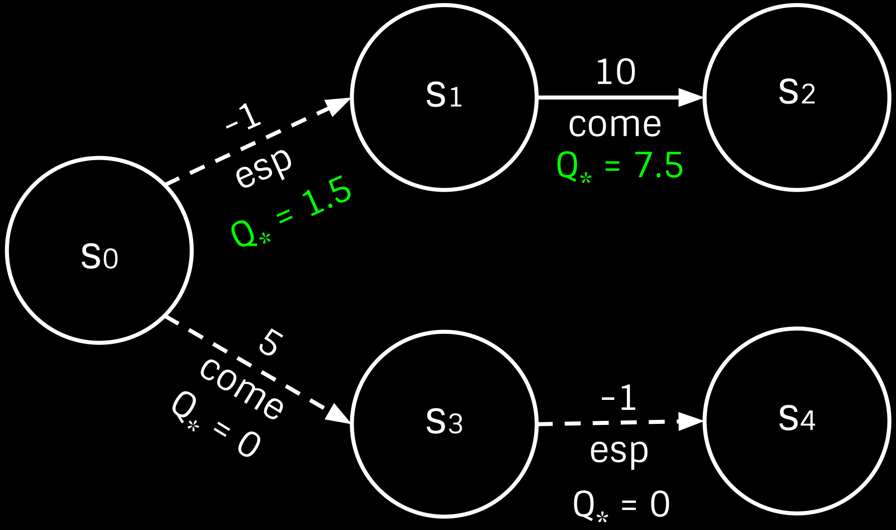
$$Q_*(s,a) \leftarrow (1-\alpha)\cdot Q_*(s,a) + \alpha \cdot Q_*^{novo}(s,a)$$

Exemplo



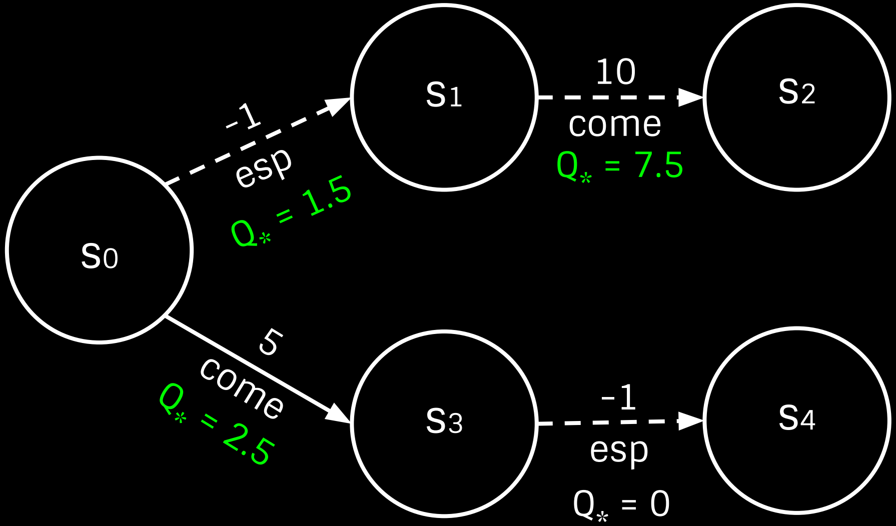
$$Q_*(s,a) \leftarrow (1-\alpha)\cdot Q_*(s,a) + \alpha \cdot Q_*^{novo}(s,a)$$

Exemplo



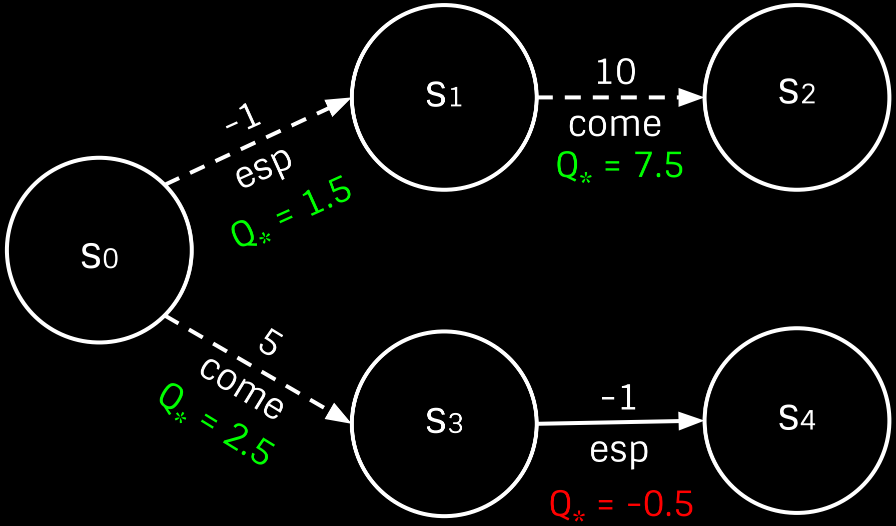
$$Q_*(s,a) \leftarrow (1-\alpha)\cdot Q_*(s,a) + \alpha \cdot Q_*^{novo}(s,a)$$

Exemplo



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

Exemplo



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

Como estimar os valores Q?

Grupo
TURING

Equação de Bellman

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

$$v_\pi(s) = E_\pi\Big[\, G_t \;\Big|\; S_t = s \,\Big]$$

$$\begin{aligned}
q_*(s,a) &= E_{\pi^*}\Big[\, G_t \;\Big|\; S_t = s, A_t = a \,\Big] \\
&= E_{\pi^*}\Big[\, R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \;\Big|\; S_t = s, A_t = a \,\Big] && \text{Retorno no instante t} \\
&= E_{\pi^*}\Big[\, R_t + \gamma\big(R_{t+1} + \gamma R_{t+2} + \cdots\big) \;\Big|\; S_t = s, A_t = a \,\Big] && \text{Isolando o } \gamma \\
&= E_{\pi^*}\Big[\, R_t + \gamma G_{t+1} \;\Big|\; S_t = s, A_t = a \,\Big] && \text{Retorno no instante t+1} \\
&= E_{\pi^*}\Big[\, R_t + \gamma v_*(S_{t+1}) \;\Big|\; S_t = s, A_t = a \,\Big] && \text{Tomando o valor médio de } G_{t+1} \\
&= E_{\pi^*}\Big[\, R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \;\Big|\; S_t = s, A_t = a \,\Big] && \text{Como o agente sempre toma a melhor ação}
\end{aligned}$$

Bootstrapping

Como calcular essa média?

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

Uma estimativa bem grosseira:
podemos considerar apenas um episódio

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

Aplicação das conclusões: Q-learning

GRUPO TURING

Q-Learning

Bootstrap gera estimativas ruins, que variam muito

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

Podemos "estabilizar" esses valores usando uma média ponderada:

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

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

Q-Learning

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

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

média ponderada

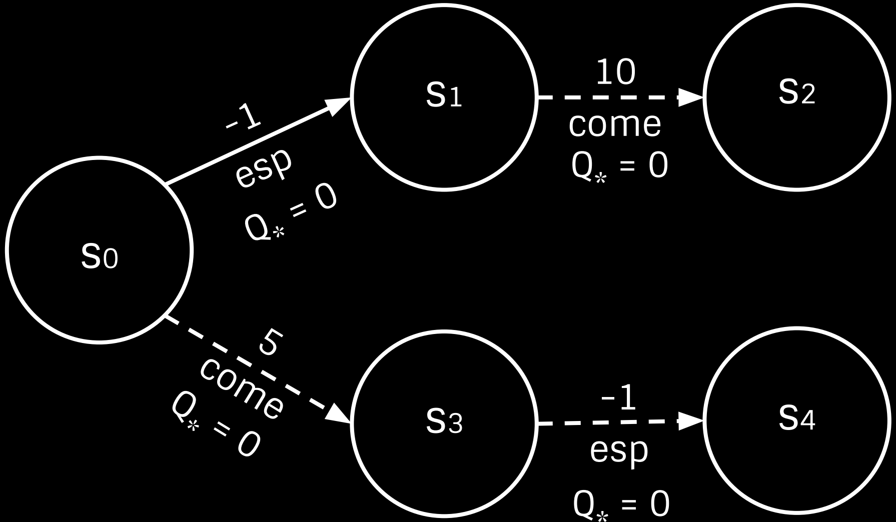
$$Q_*(s,a) \leftarrow \underbrace{Q_*(s,a)}_{\substack{\text{valor}\\\text{antigo}}} + \underbrace{\alpha}_{\substack{\text{taxa de}\\\text{aprendizado}}} \cdot \overbrace{\left(\underbrace{r + \gamma \max_{a'} Q_*(s',a')}_{Q_{\text{bootstrap}}(s,a)} - Q_*(s,a)\right)}^{\text{erro}}$$

minimização do erro

Exemplo (ep. 1)

$\alpha = 0.5$

$\gamma = 0.9$



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

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

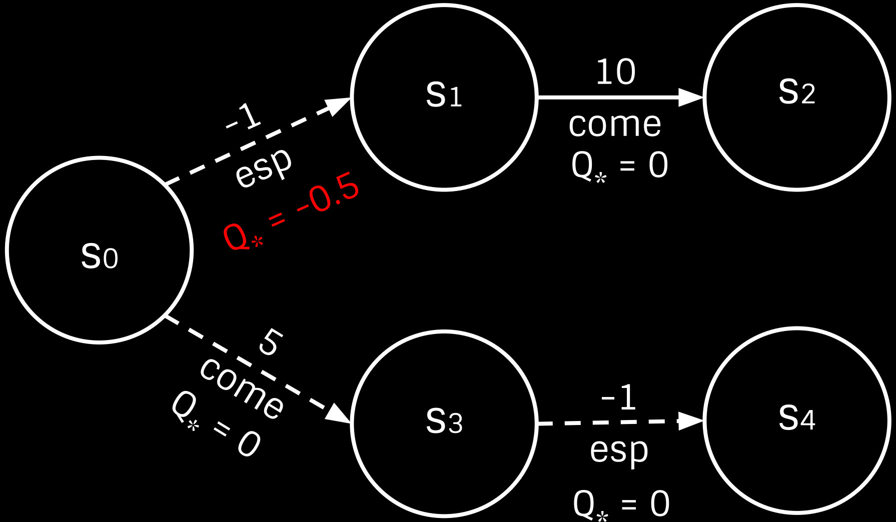
$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot 0 + 0.5 \cdot (-1)$$

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

Exemplo (ep. 1)

$\alpha = 0.5$

$\gamma = 0.9$



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

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

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

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

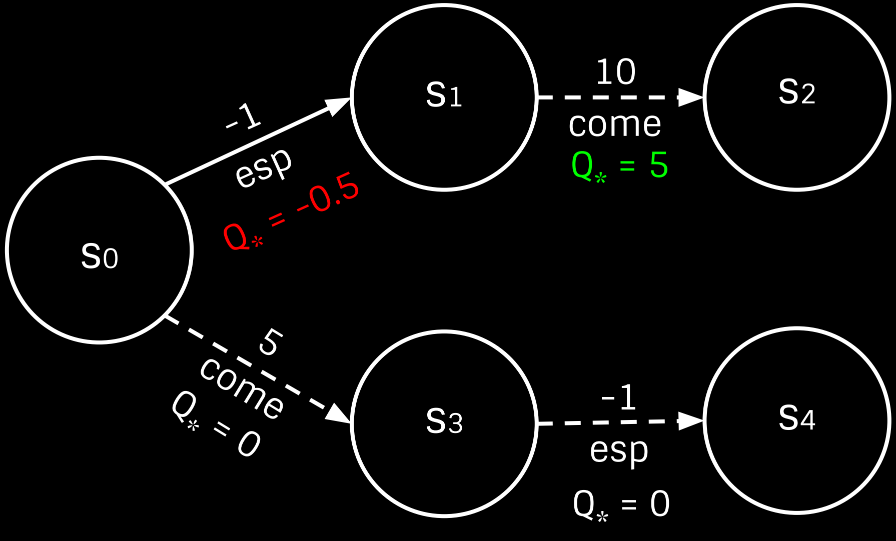
$$Q_{boot}(s_1, \text{come}) = r + \gamma \max_{a'} Q_*(s_2, a') = 10 + 0.9 \cdot 0 = 10$$

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

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

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

Exemplo (ep. 2)



$\alpha = 0.5$

$\gamma = 0.9$

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

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

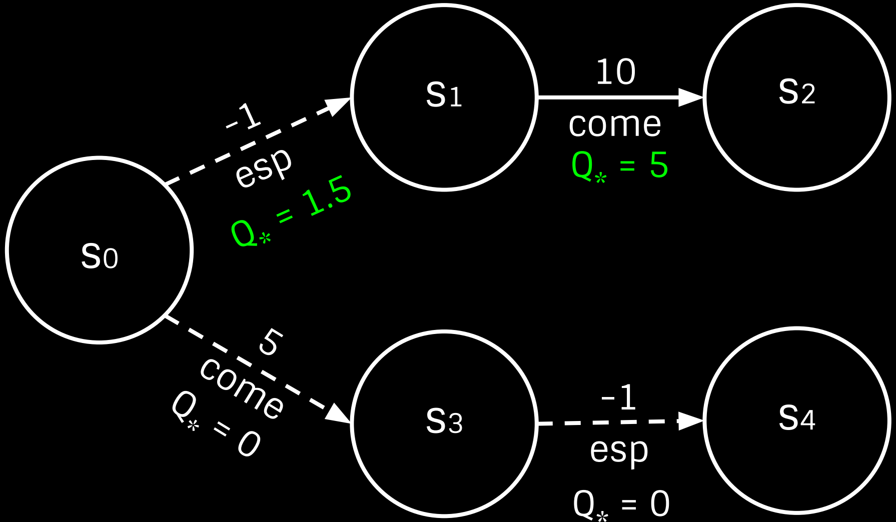
$$Q_*(s_0, \text{esp}) \leftarrow 0.5 \cdot (-0.5) + 0.5 \cdot 3.5$$

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

Exemplo (ep. 2)

$\alpha = 0.5$

$\gamma = 0.9$



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

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

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

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

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

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

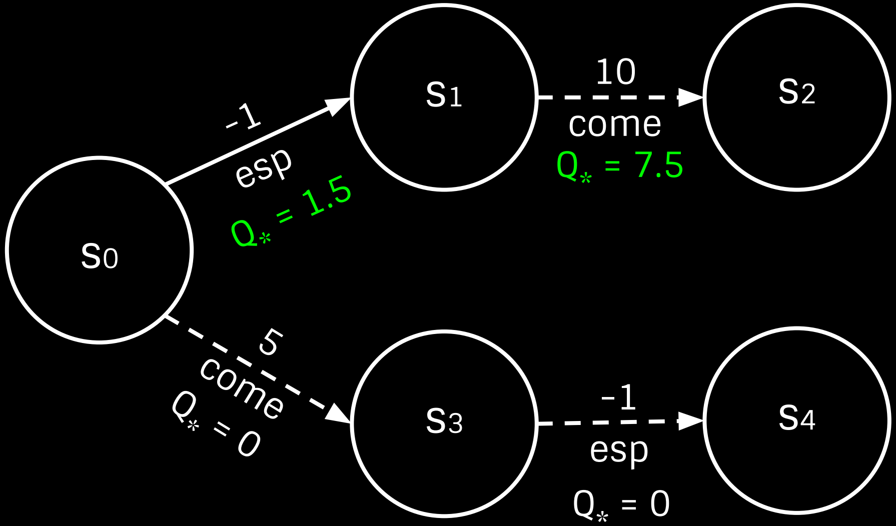
$$Q_*(s_1, \text{come}) \leftarrow 0.5 \cdot 5 + 0.5 \cdot 10$$

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

Exemplo (ep. 3)

$\alpha = 0.5$

$\gamma = 0.9$



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

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

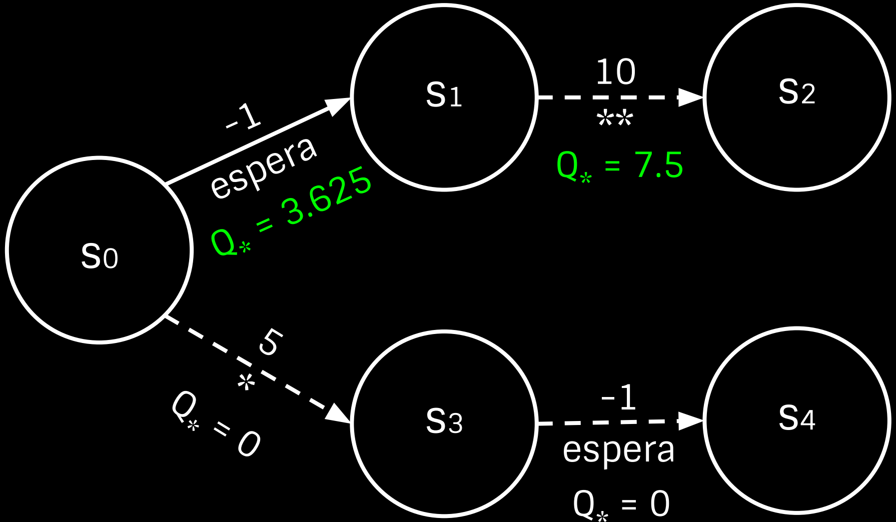
$$Q_*(s_0, \text{espera}) \leftarrow 0.5 \cdot 1.5 + 0.5 \cdot 5.75$$

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

Exemplo (ep. 3)

$\alpha = 0.5$

$\gamma = 0.9$



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

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

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

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

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

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

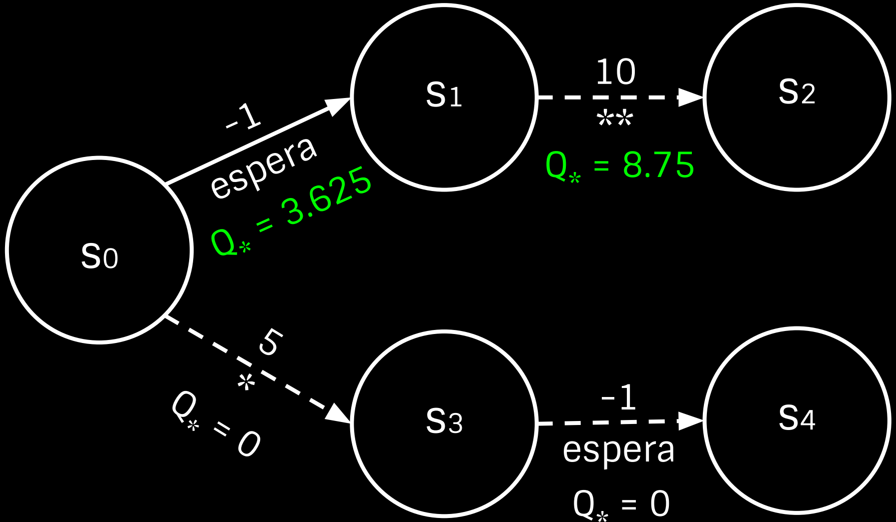
$$Q_*(s_1, \text{come}) \leftarrow 0.5 \cdot 7.5 + 0.5 \cdot 10$$

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

Exemplo (ep. 3)

$\alpha = 0.5$

$\gamma = 0.9$



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 $\alpha, \gamma \in (0, 1]$, ε pequeno > 0.

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

Loop para cada episódio:

    Inicialize S

    Loop para cada instante do episódio:

        Escolha A usando uma política derivada de Q (e.g, ε-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)

Livro:

- Reinforcement Learning: An Introduction (Sutton, Barto)