Markov Decision Process

MDP

Definição de MDP

Processo de Controle estocástico em tempo discreto

Baseado em tomada de decisão

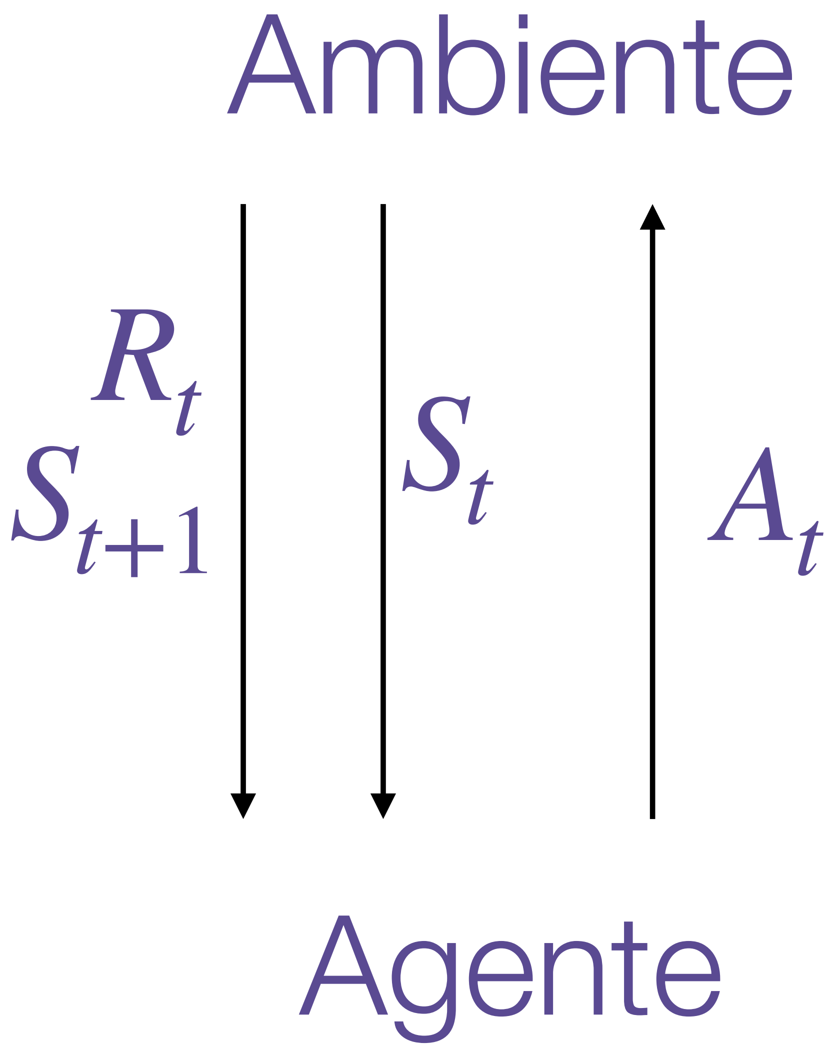
Intervalo de tempo finitos

Estados futuros dependem

Parcialmente das acoes do agente

Definição de MDP

Generalização



Definição de MDP

Nao se esqueça:

$$(S, A, R, P)$$

Conjunto de todos possíveis estados

Conjunto de acoes Tomadas em cada Estado

Conjunto de Pagamentos para Cada par (s, a)

Probabilidades de passar De um estado para outro Tomando uma ação

Definição de MDP

Memória

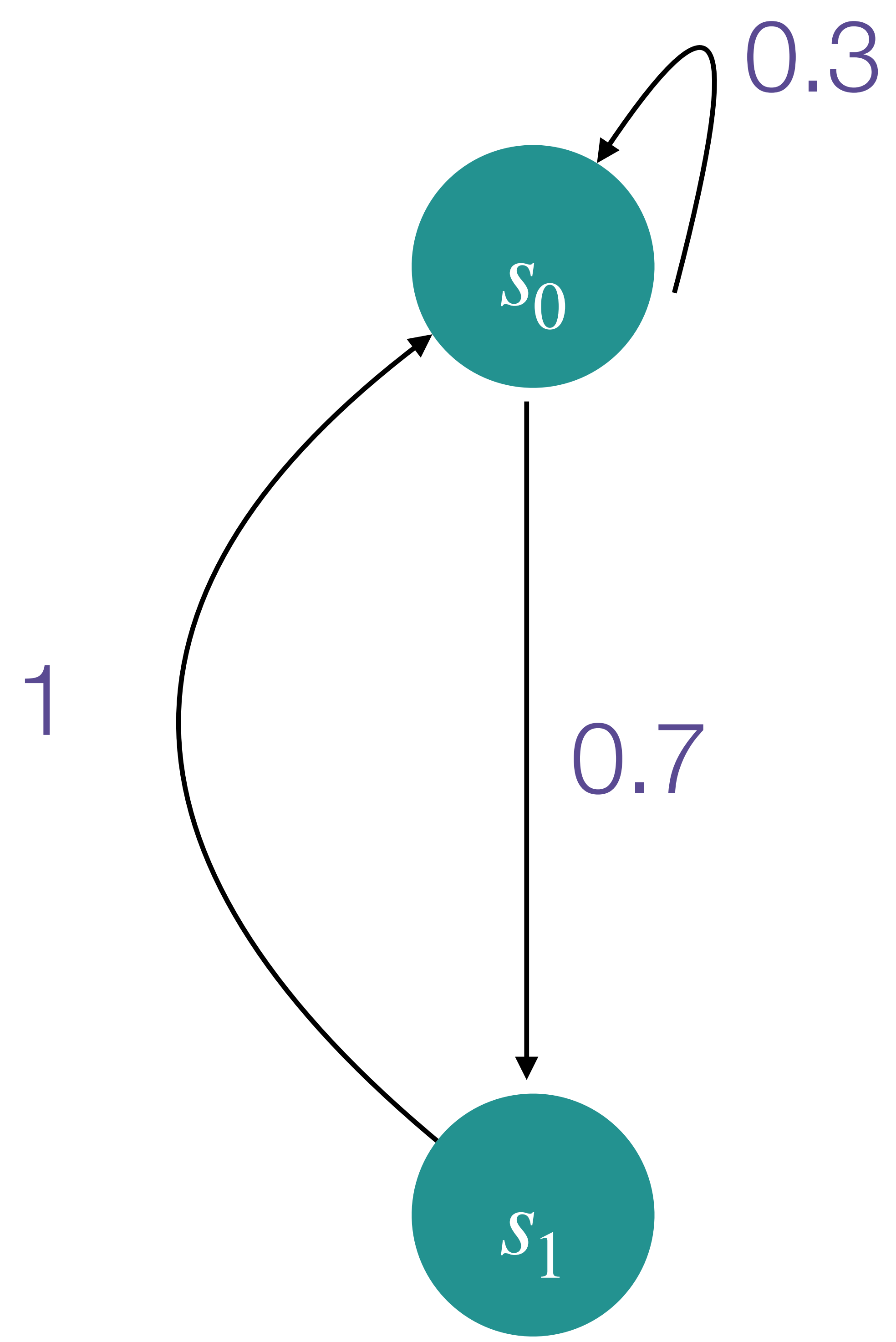
- MDP não possui memória (*memoryless*)

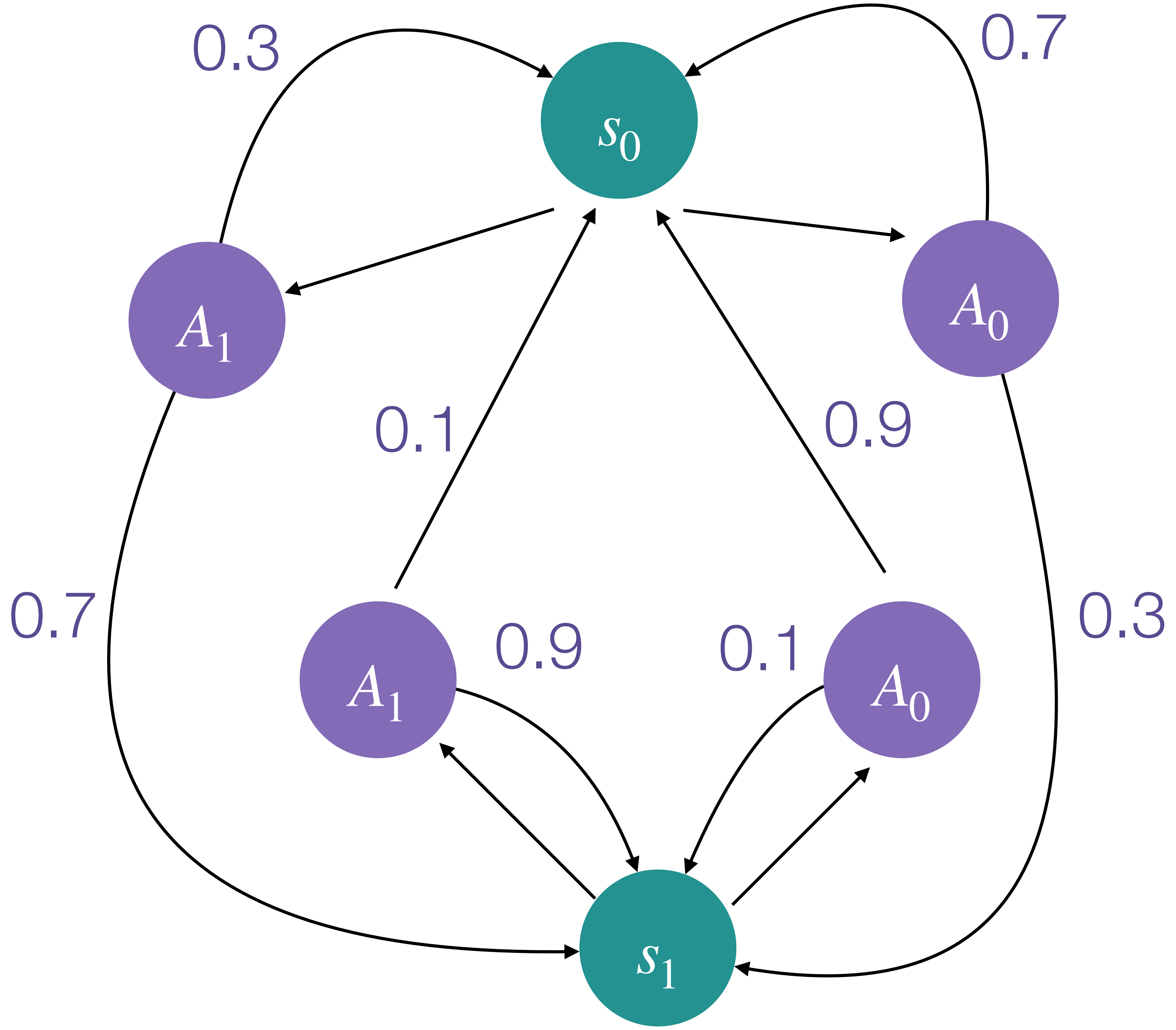
$$P[S_{t+1} \mid S = s_t] = P[S_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0]$$

O próximo estado só
Depende do atual estado

Definição de MDP

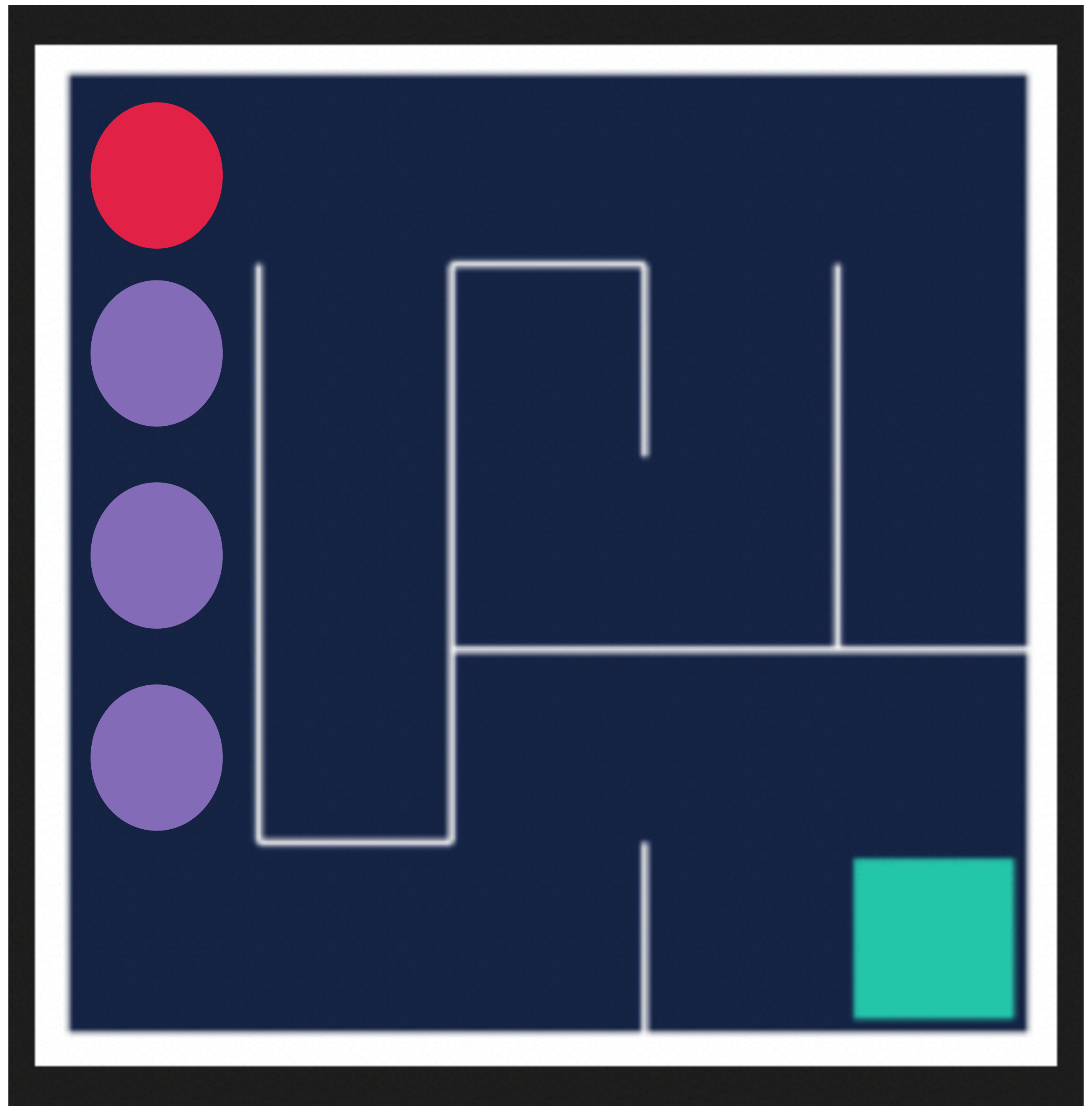
Extensão de uma cadeia de Markov





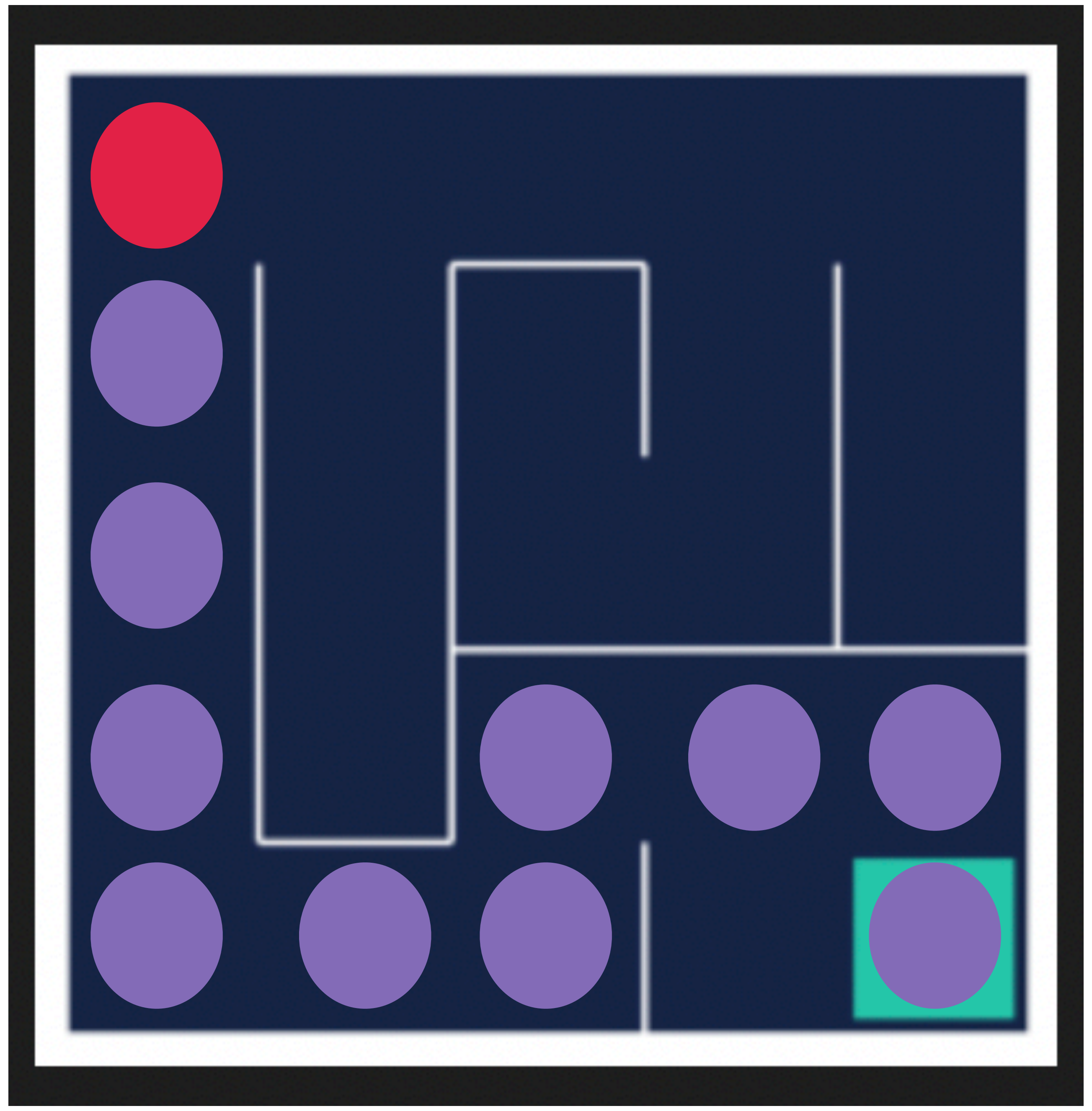
Trajetória vs. Episódio

$$\tau = S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3$$



Trajetória vs. Episódio

$$\tau = S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots, R_T, S_T$$



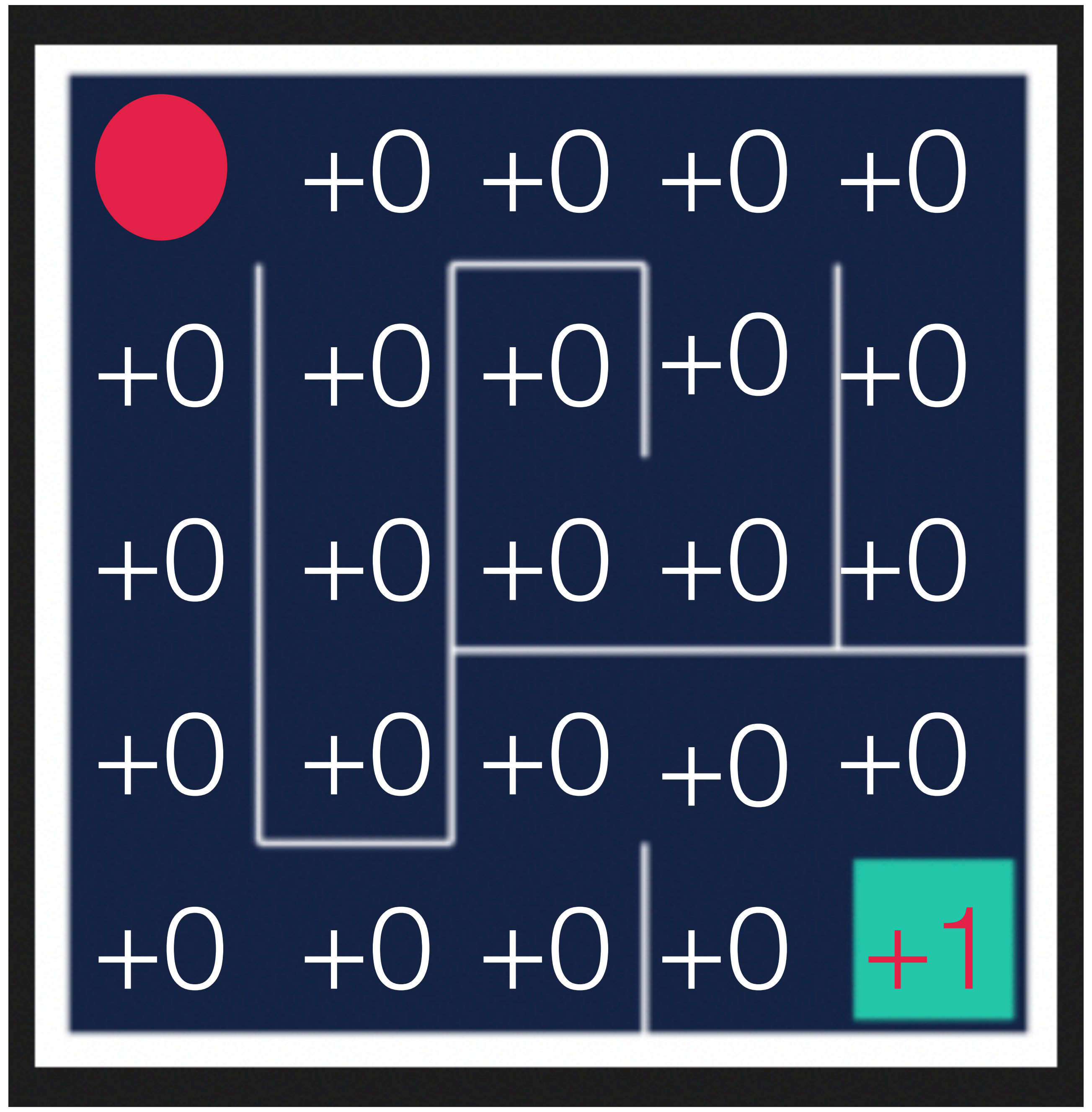
Pagamento vs. Retorno

- O objetivo de qualquer tarefa de RL é de maximizar a soma dos pagamentos R_t
- Um pagamento grande a curto prazo pode ser piorar resultados a longo prazo
- Retorno G_t é a soma a longo prazo dos pagamentos R_t

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$$

Fator de Desconto

- Qual o incentivo para tomar o caminho mais curto?
- Pensemos no retorno a longo prazo

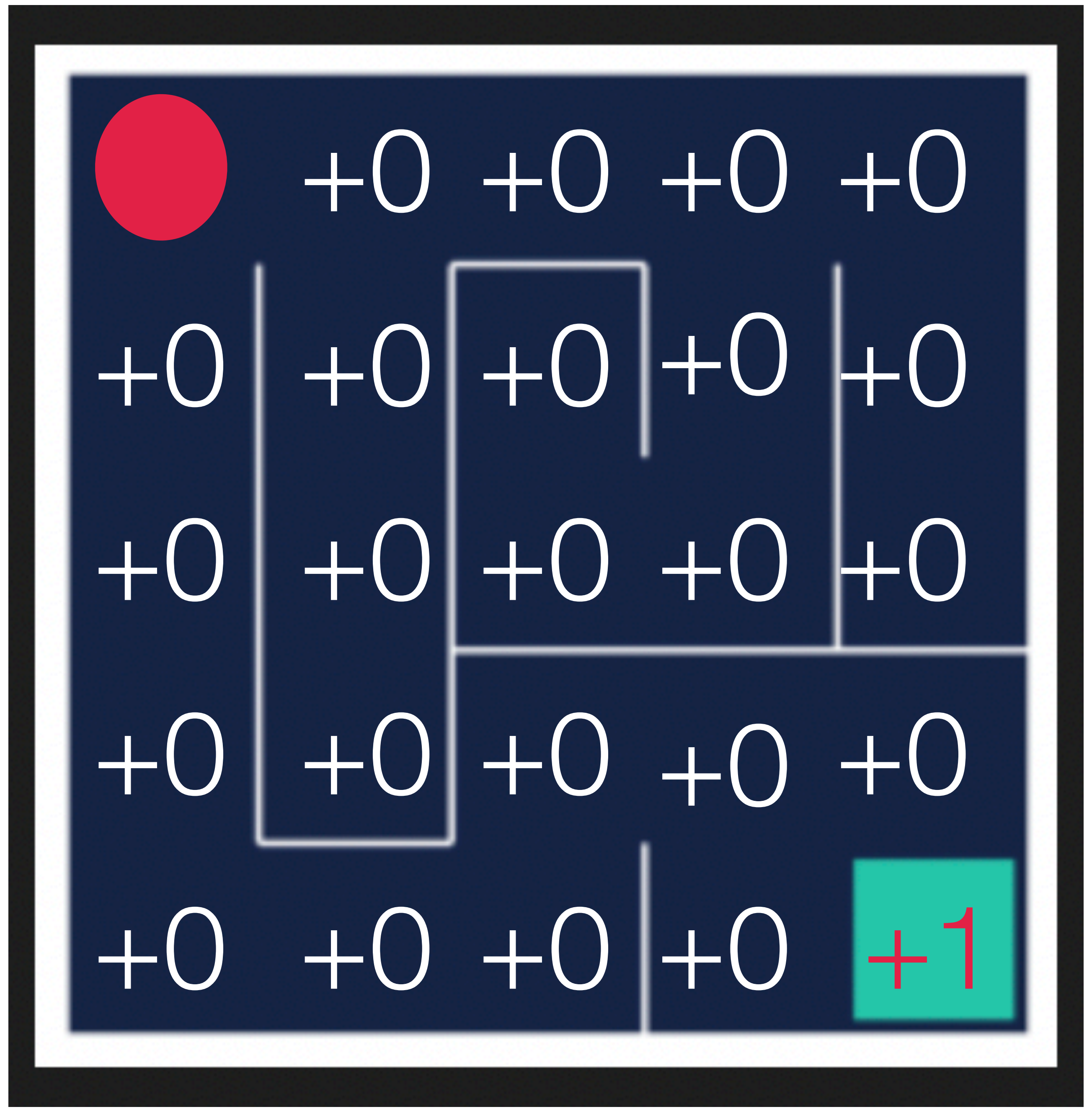


Fator de Desconto

$$G_0 = R_1 + \gamma R_2 + \gamma R_3 + \ldots + \gamma^{T-t-1} R_T \qquad \gamma \in [0,1]$$

$$G_0 = R_1 + \sum_{i=2}^{T-t-1} R_i$$

- E se γ for 0?
- E se γ for 1?



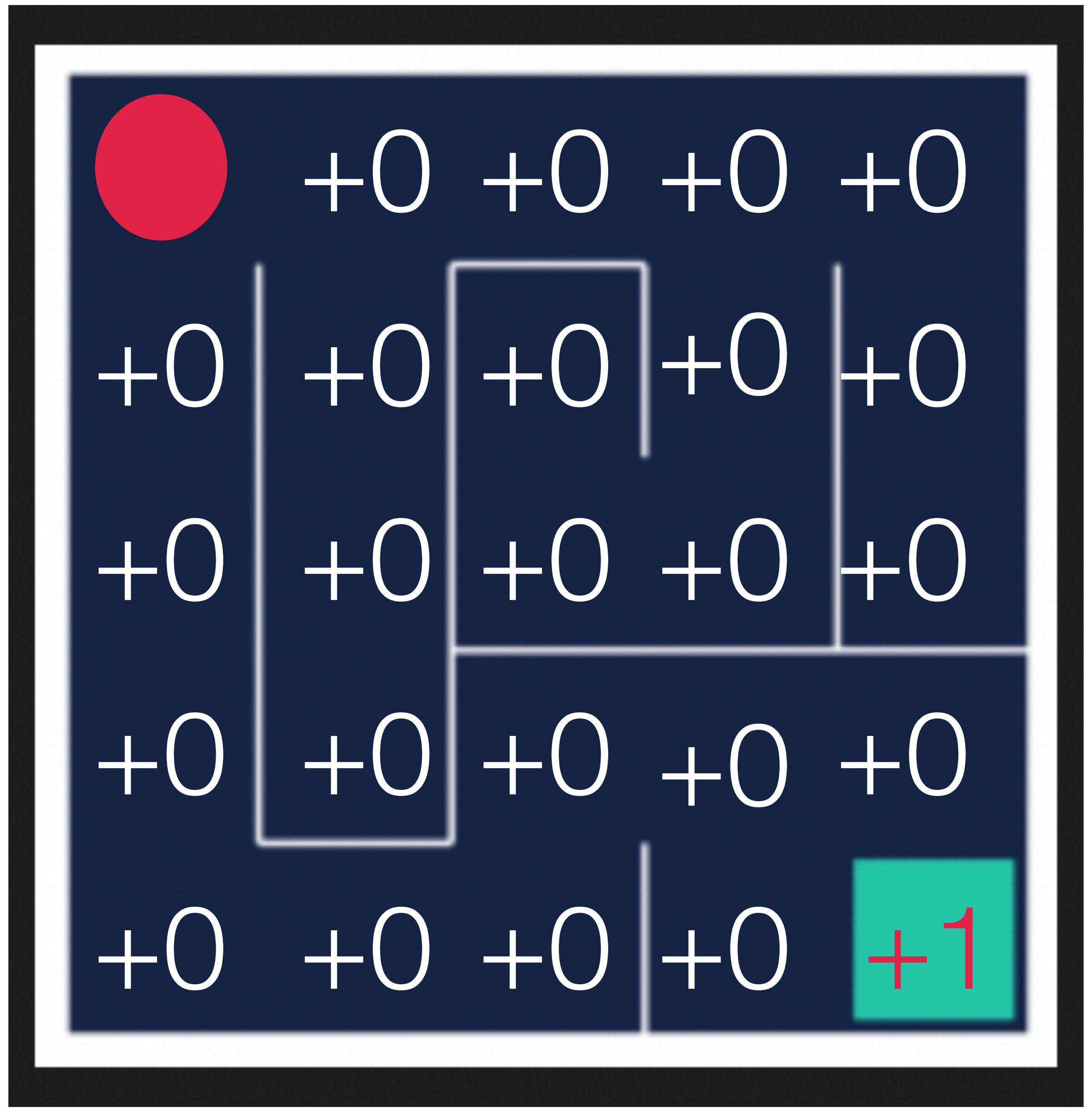
Fator de Desconto

$$G_0 = R_1 + \gamma R_2 + \gamma R_3 + \ldots + \gamma^{T-t-1} R_T \qquad \gamma \in [0,1]$$

$$G_0 = R_1 + \sum_{i=2}^{T-t-1} R_i$$

- E se γ for 0?
- E se γ for 1?

Objetivo: Maximizar a soma dos pagamentos **descontados**



Políticas

- Objetivo no MDP é encontrar uma politica ótima $\pi^\star$ que maximize G, ou seja, um conjunto de ações que maximizem G
- Lembre-se que cada politica π traz um resultado G diferente

Valores de estado e ações

- Como se avaliar uma politica? Valores de estado

Valor obtido começando do estado s

Interagindo com o ambiente seguindo π

Esperança

Até o final do episódio

$$v_\pi(s) = \mathbb{E}[G_t \,|\, S_t = s]$$

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T \,|\, S_t = s]$$

Valores de estado e ações

- Avaliando pelos pares de ações e estados

Valor obtido começando do estado s, tomando ação a

Interagindo com o ambiente seguindo π

$$q_\pi(s, a) = \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

$$q_\pi(s, a) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T \mid S_t = s, A_t = a]$$

Equações de Bellman

- Como achar o melhor π tomando por base os valores $v(s)$

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t \mid S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T \mid S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \sum_a \pi(a \mid s) \sum_{s',r} p(s, r \mid s, a)[r + \gamma v_\pi(s')]
\end{aligned}
$$

A esperança pode ser reescrita como a probalidade de tomar uma ação seguindo a politica π multiplicado
Pelo retorno obtido por tomar a ação a

Equações de Bellman

- Para valores $q(s, a)$

$$
\begin{aligned}
q_\pi(s, a) &= \mathbb{E}[G_t \mid S_t = s, A_t = a] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T \mid S_t = s, A_t = a] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \sum_{s', r} p(s', r \mid s, a)[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a')]
\end{aligned}
$$

Probabilidade de alcançar cada estado sucessor s' sabendo que se tomou acao a multiplicado pelo pagamento r imediato somado dos valores descontados $q(s, a)$ de cada ação a' no estado sucessor s' ponderado pela probabilidade $P(a' \mid s')$ de tomar a ação a da politica π

Resolvendo um MDP

- Uma politica ótima π^* escolhe um conjunto de ações A^*de forma a maximizar $v(s)$ou $q(s, a)$

$$\pi_*(s) = \text{argmax}_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v^*(s)]$$

$$\pi_*(s) = \text{argmax}_a q^*(s, a)$$

Resolvendo um MDP

- Há um pequeno porém:

Para se achar π^*, deve-se
Saber os valores ótimos v^* ou q^*

Para se achar os valores ótimos
v^* ou q^* deve-se saber π^*

Resolvendo um MDP

- Voltamos as equações, de Bellman e as resolvemos iterativamente

$$v*(s) = \mathrm{max}_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v*(s')]$$

$$q*(s, a) = \sum_{s',r} p(s', r \mid s, a)[r + \gamma \mathrm{max}_{a'} q*(s', a')]$$

Resolvendo um MDP

Abra o Notebook MDP.ipynb