

Markov Decision Process

MDP

Definição de MDP

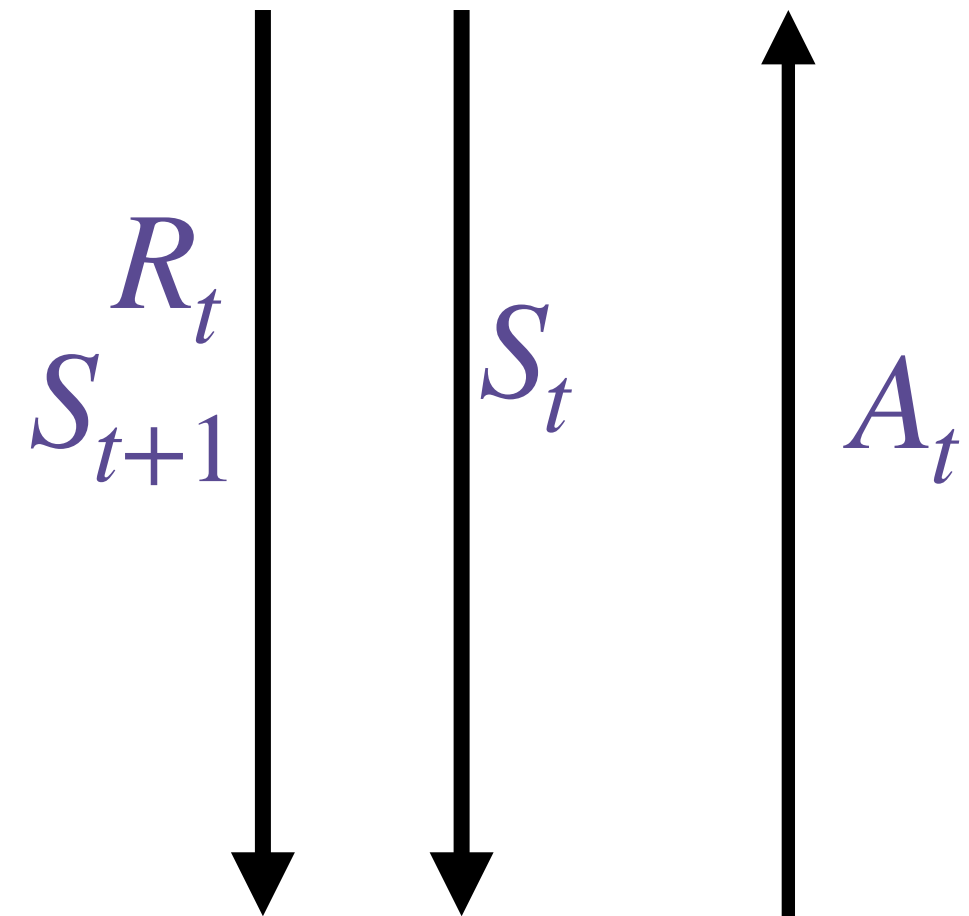
Processo de Controle estocástico em tempo discreto

- Baseado em tomada de decisão
- Intervalo de tempo finitos
- Estados futuros dependem parcialmente das ações do agente

Definição de MDP

Generalização

Ambiente



Agente

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

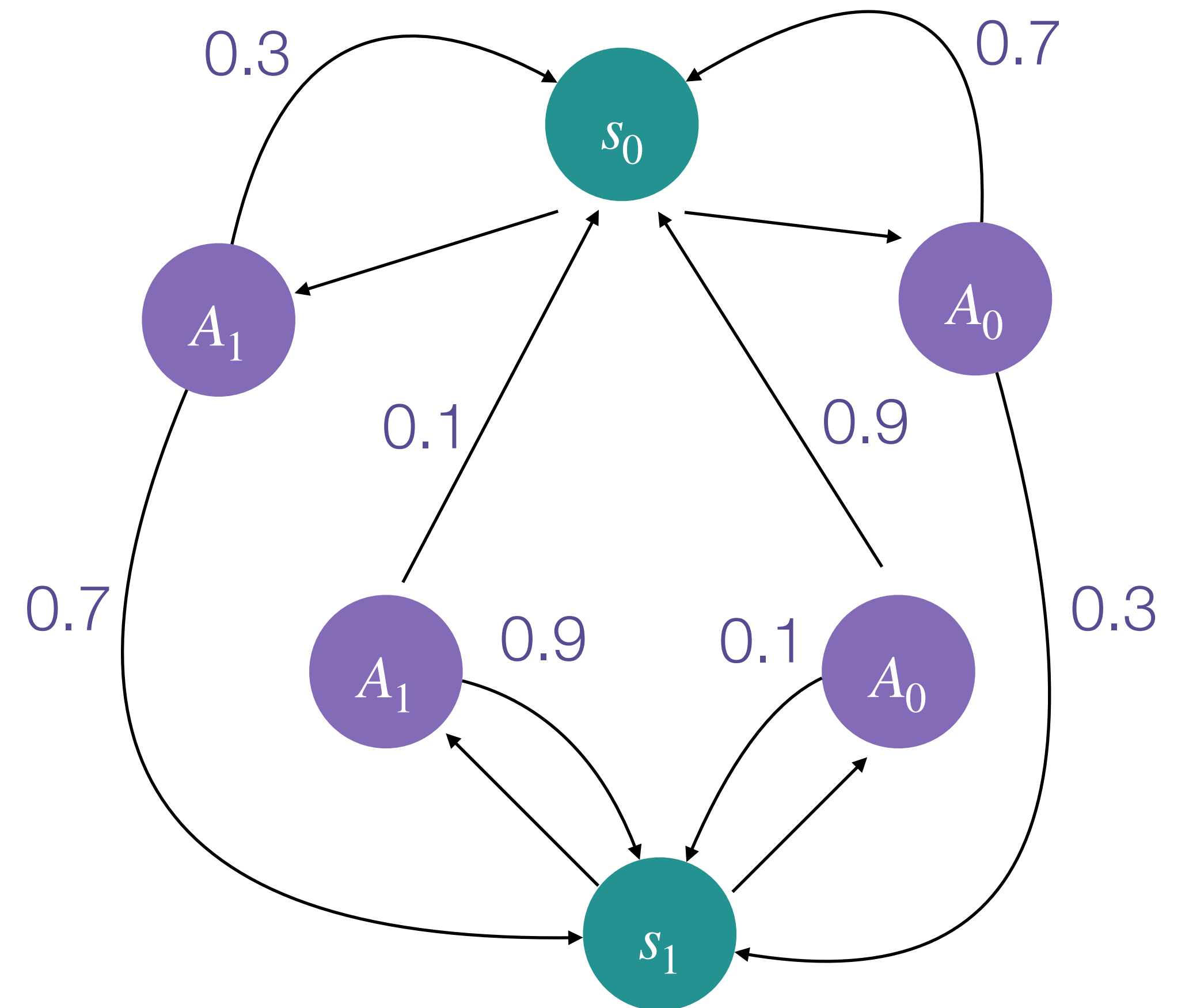
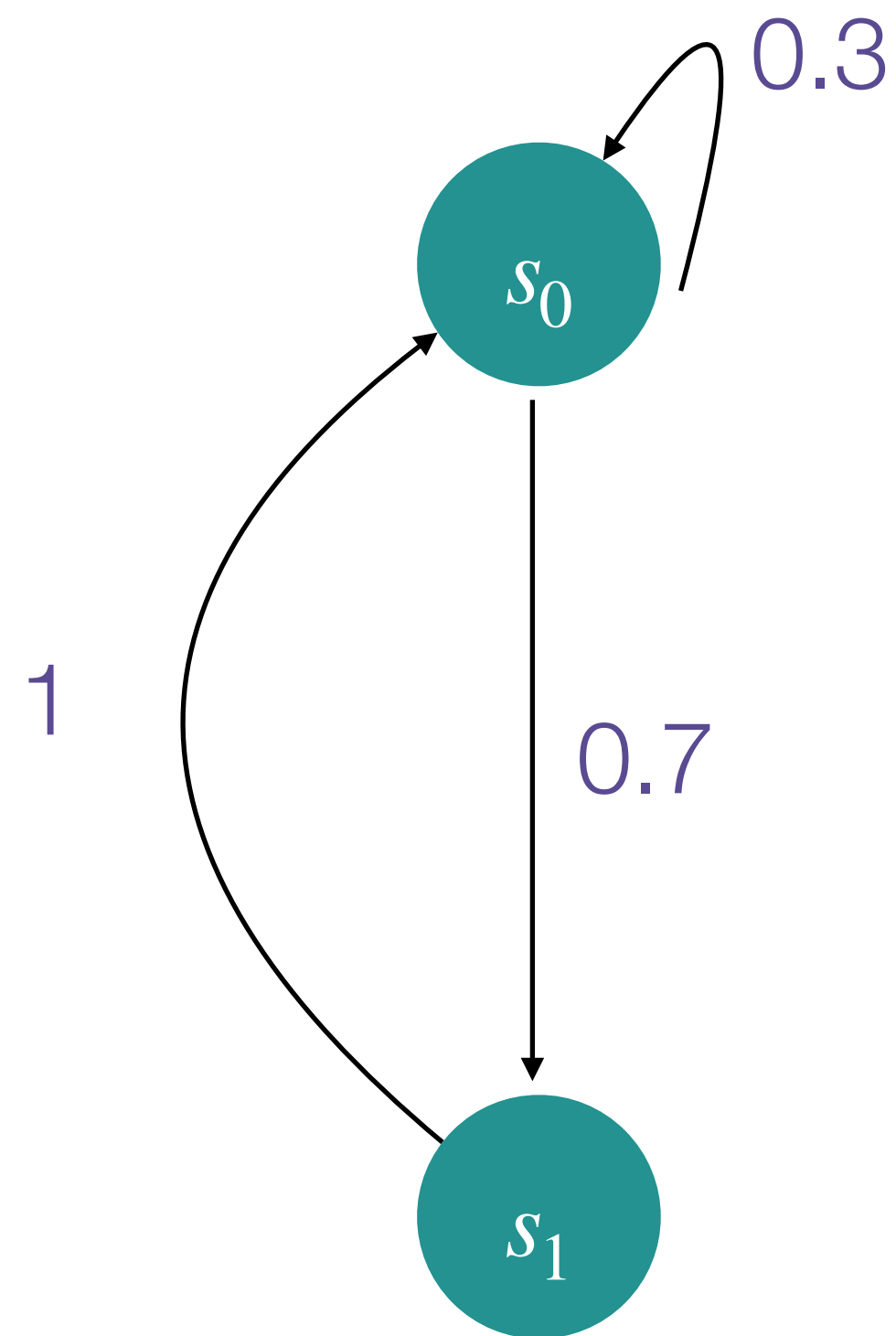
- MDP não possui memória (*memoryless*)

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

O próximo estado só
Depende do atual estado

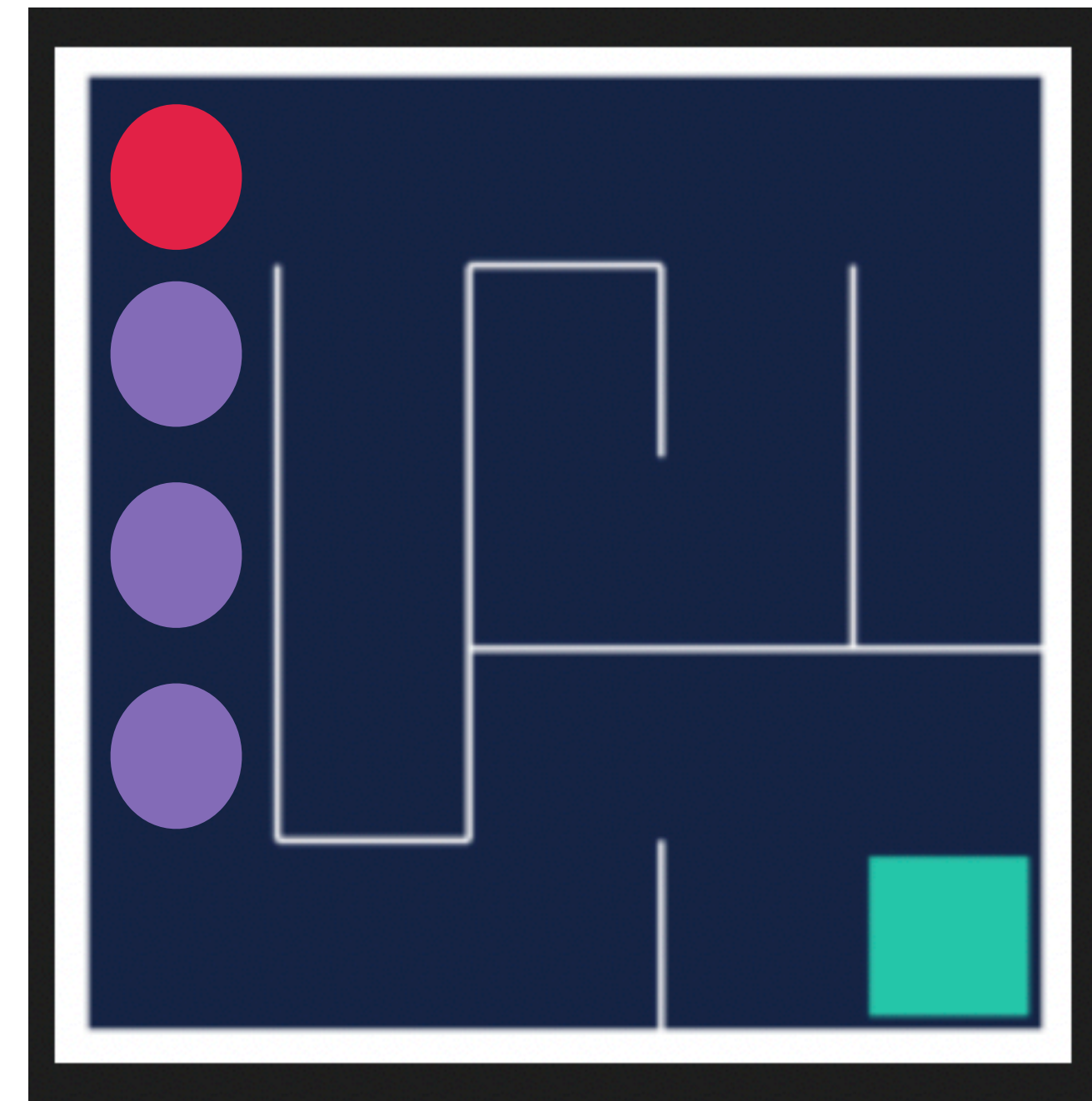
Definição de MDP

Extensão de uma cadeia de Markov



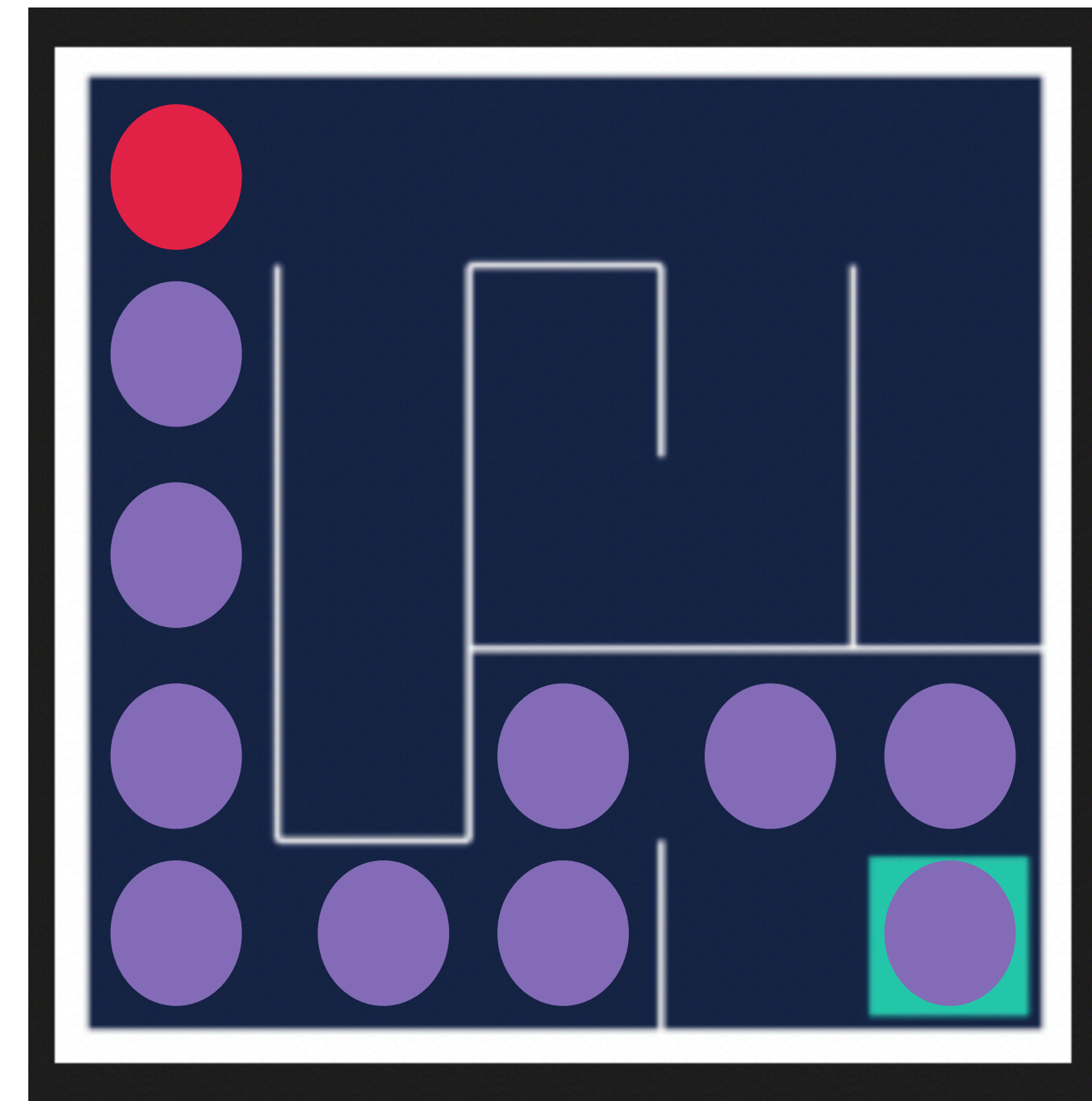
Trajetória vs. Episódio

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



Trajetória vs. Episódio

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



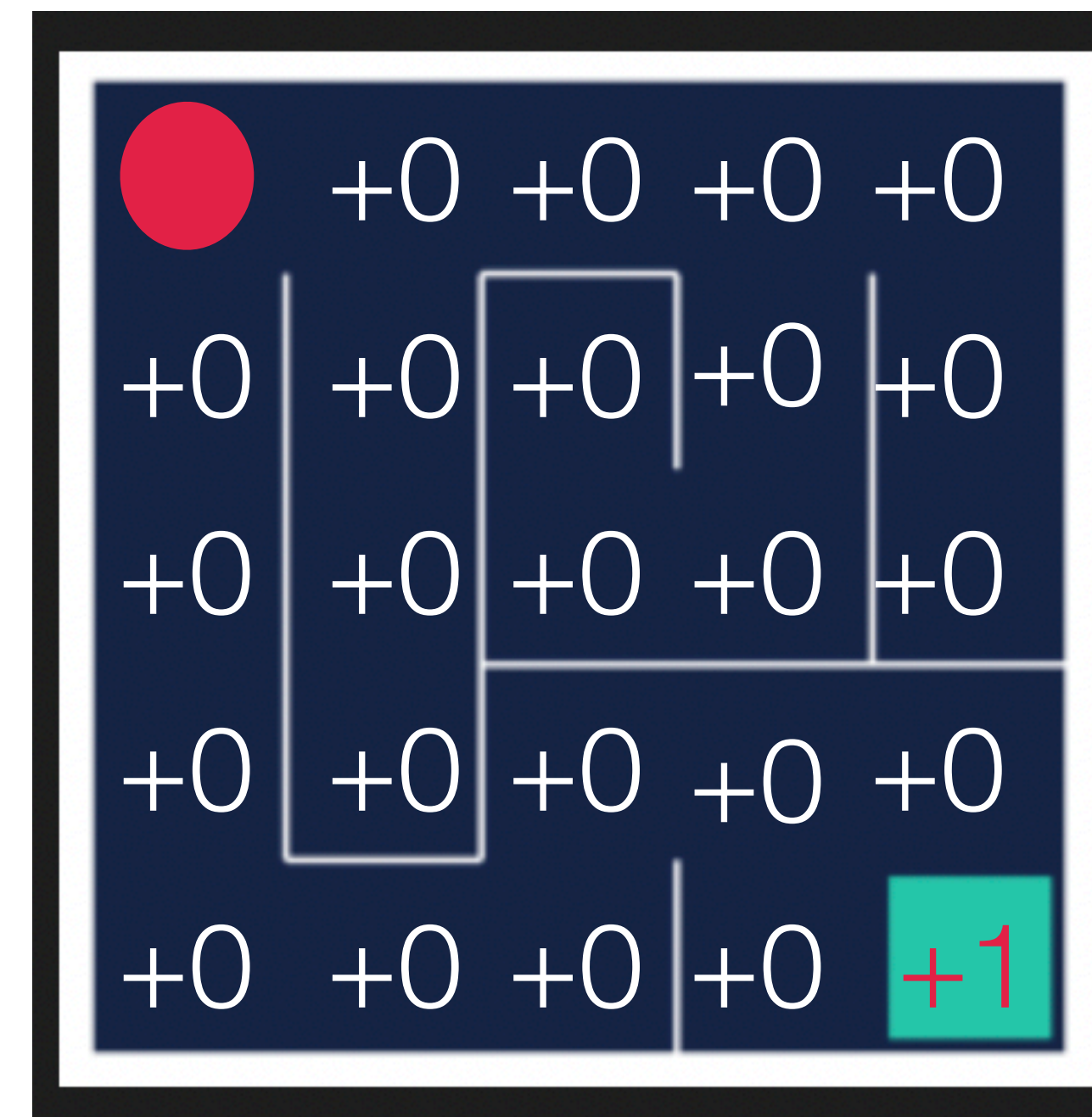
Pagamento vs. Retorno

- O objetivo de qualquer tarefa de RL é de maximizar a soma dos pagamentos R_t
- Um pagamento grande a curto prazo pode 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



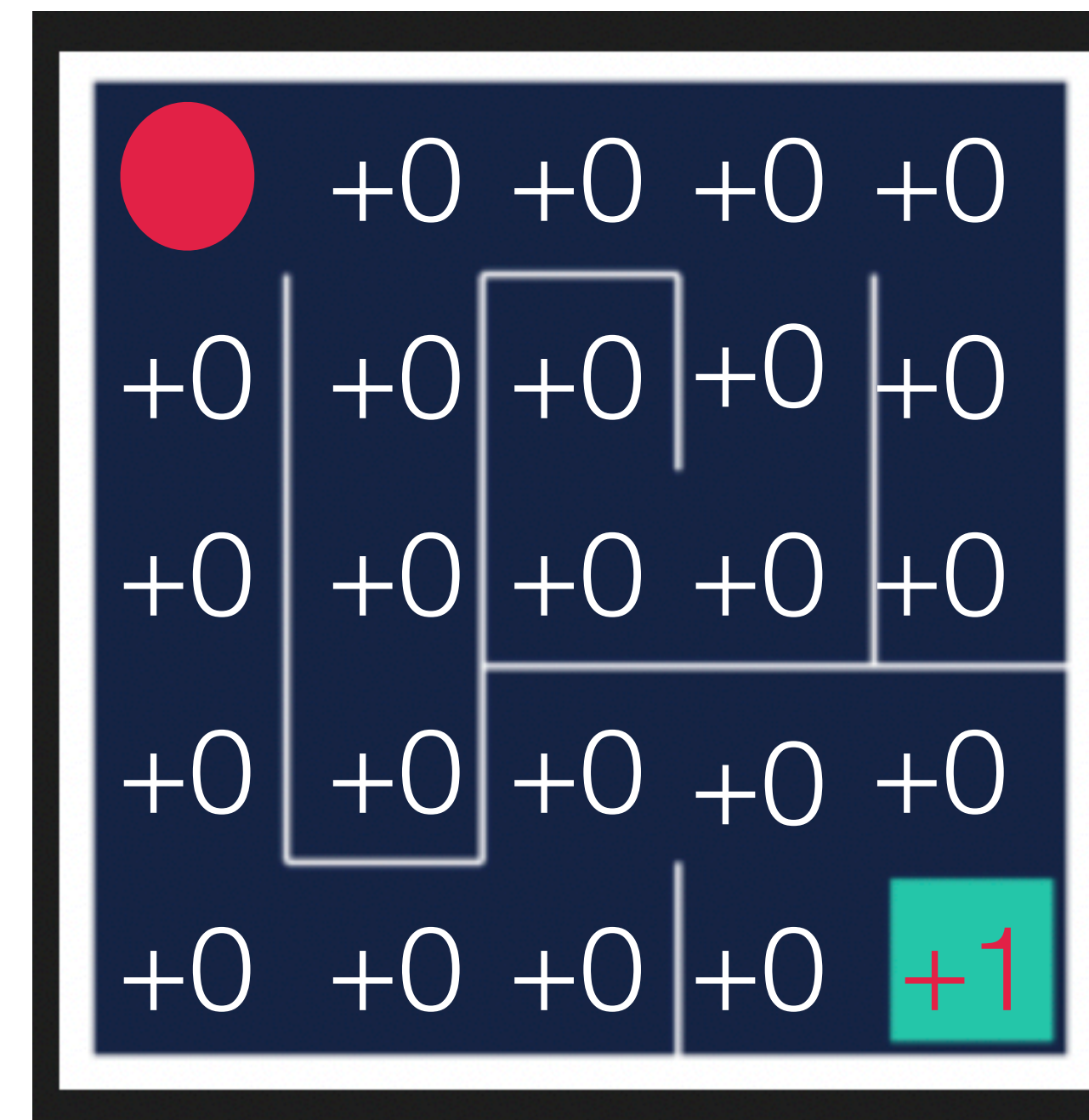
Fator de Desconto

$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \dots + \gamma^{T-t-1} R_T \quad \gamma \in [0,1]$$

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

- E se γ for 0?
- E se γ for 1?

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



Políticas

- Objetivo no MDP é encontrar uma política ótima π^\star que maximize G , ou seja, um conjunto de ações que maximizem G
- Lembre-se que cada política 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

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

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

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


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

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

Equações de Bellman

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

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

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

Equações de Bellman

- Para valores $q(s, a)$

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

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

Resolvendo um MDP

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

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

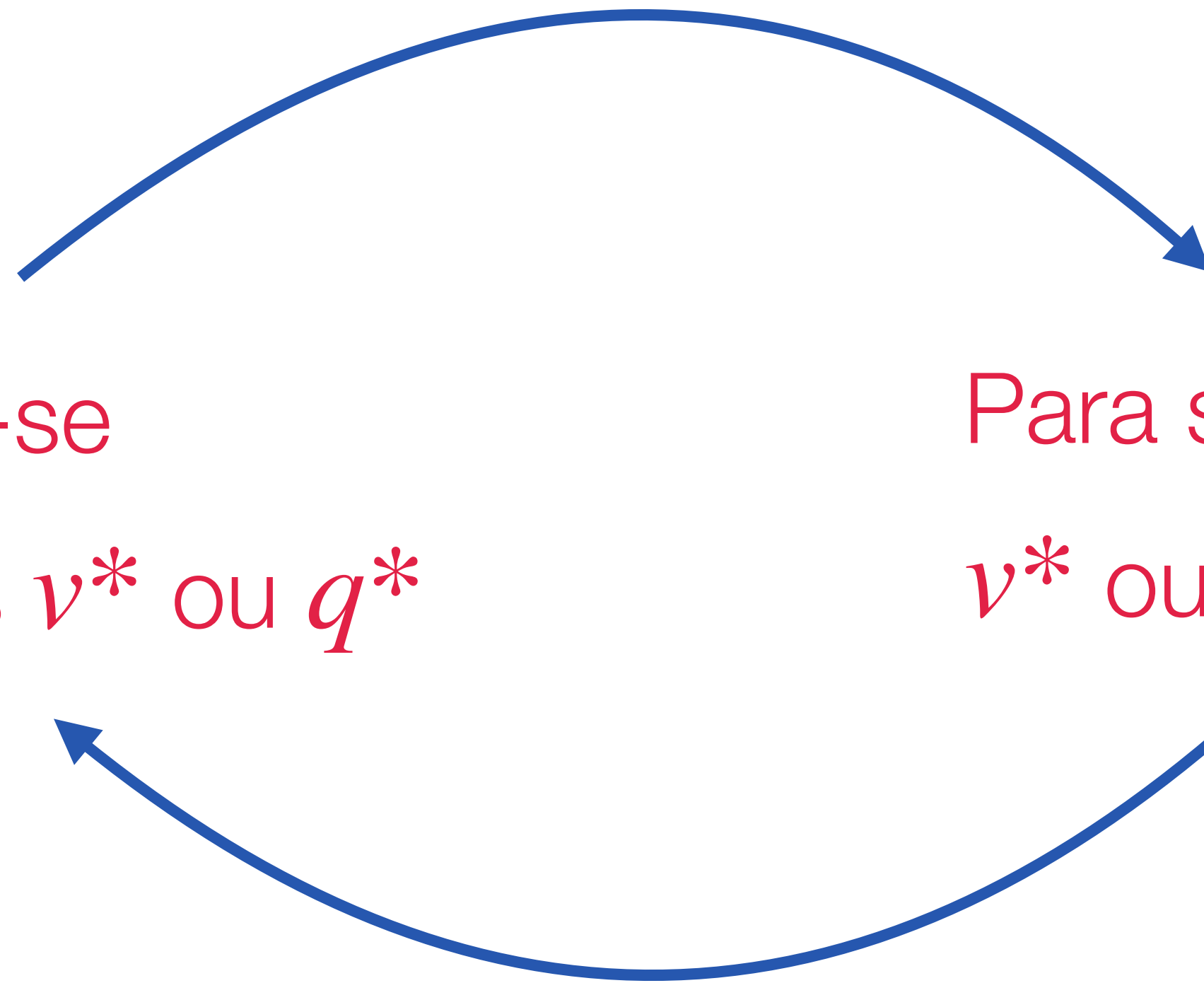
$$\pi_*(s) = \operatorname{argmax}_a q^*(s, a)$$

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

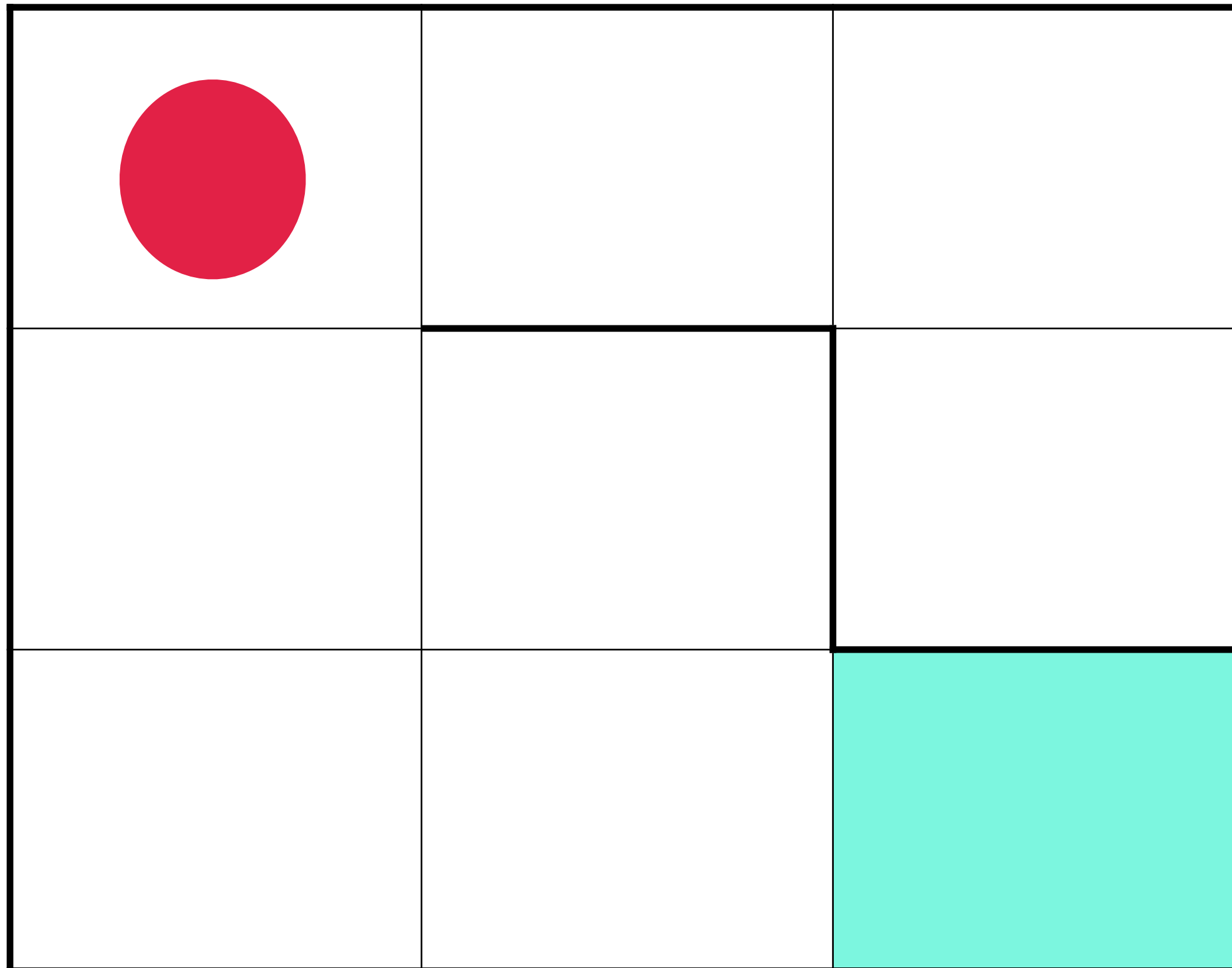
- Voltamos as equações, de Bellman e as resolvemos iterativamente

$$v^*(s) = \max_a \sum_{s',r} p(s', r | s, a) [r + \gamma v^*(s')]$$

$$q^*(s, a) = \sum_{s',r} p(s', r | s, a) [r + \gamma \max_{a'} q^*(s', a')]$$

Exemplo visual

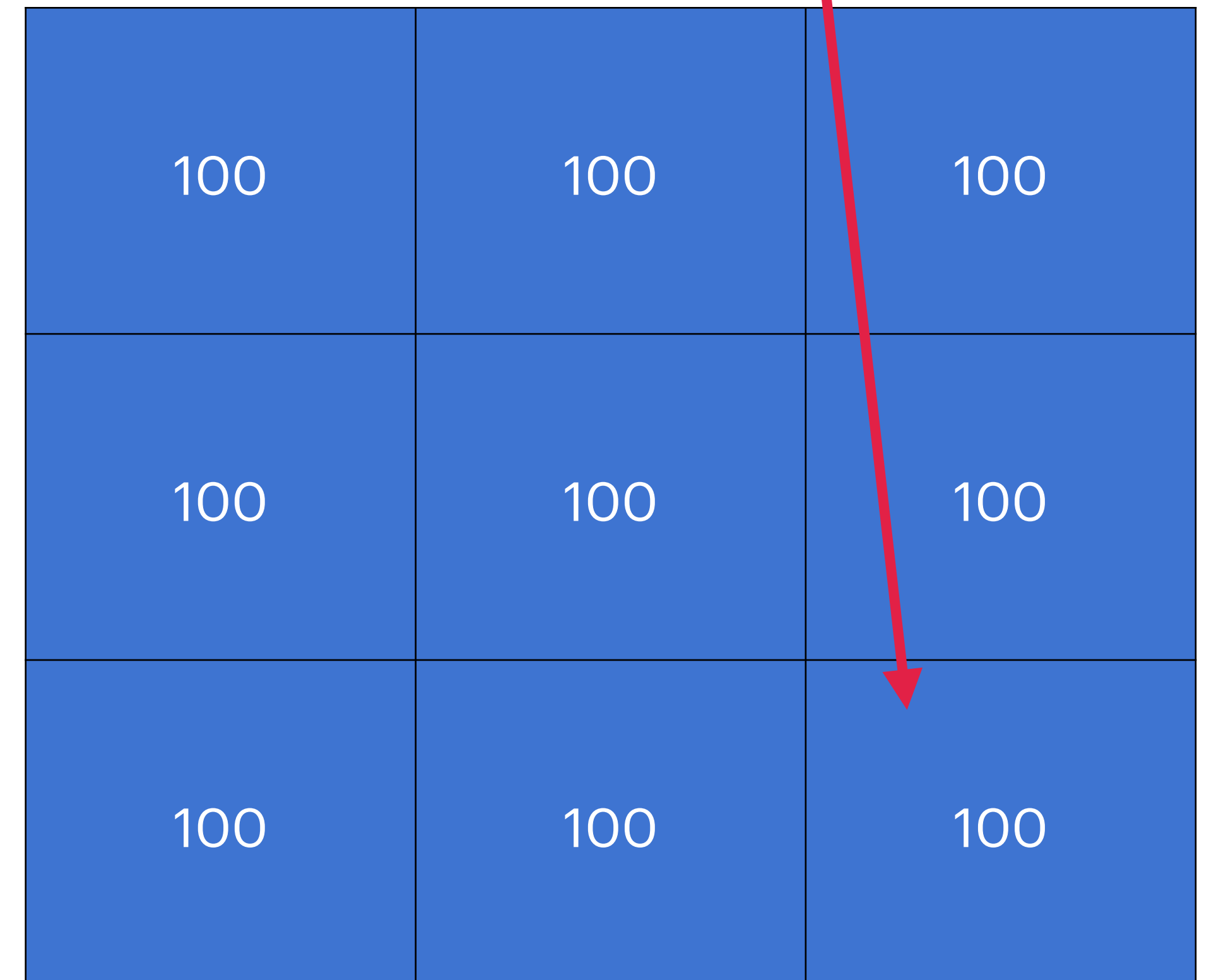
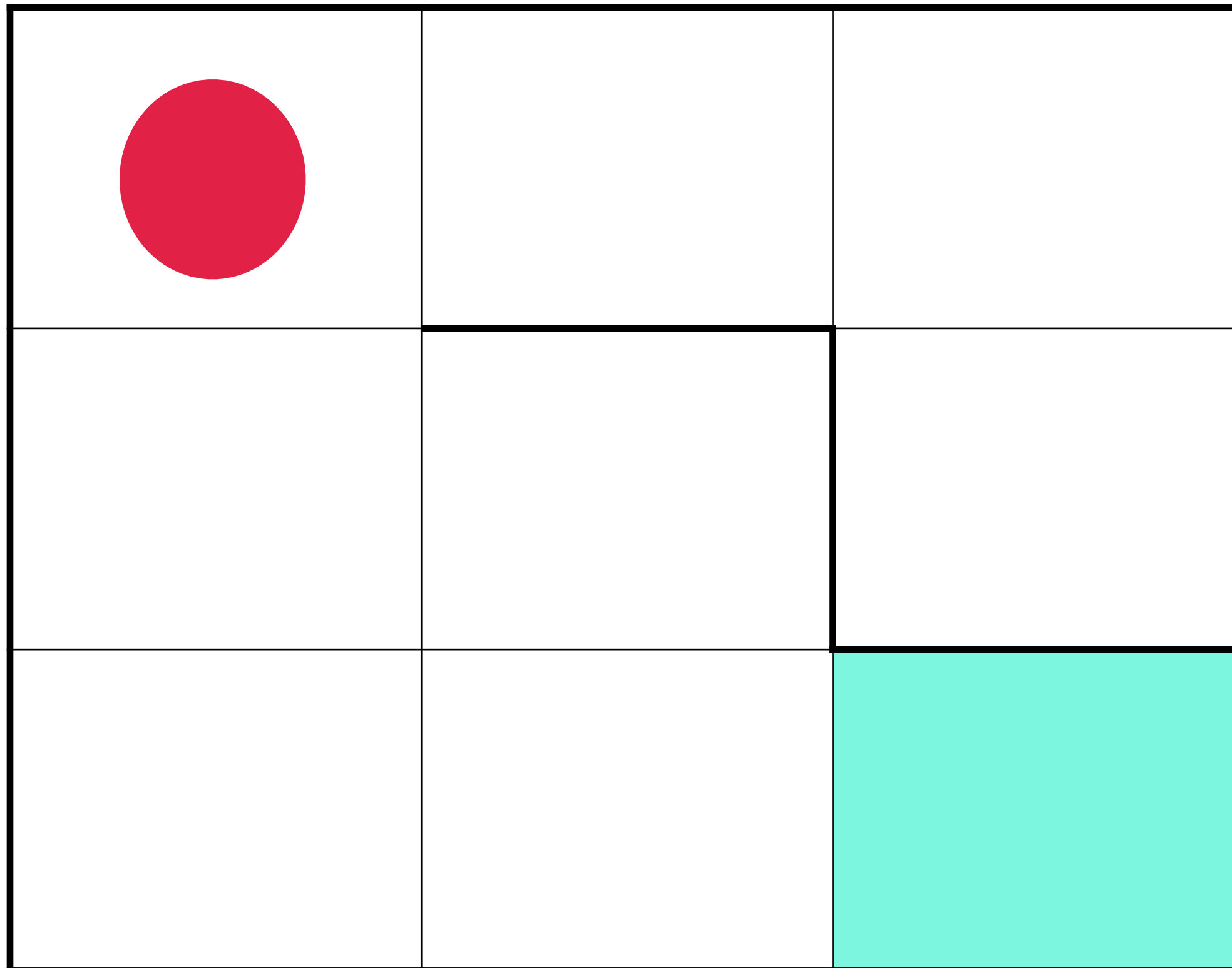
- Consideremos esse pequeno labirinto 3x3



Exemplo visual

- Para valores $v(s)$, $t = 1$

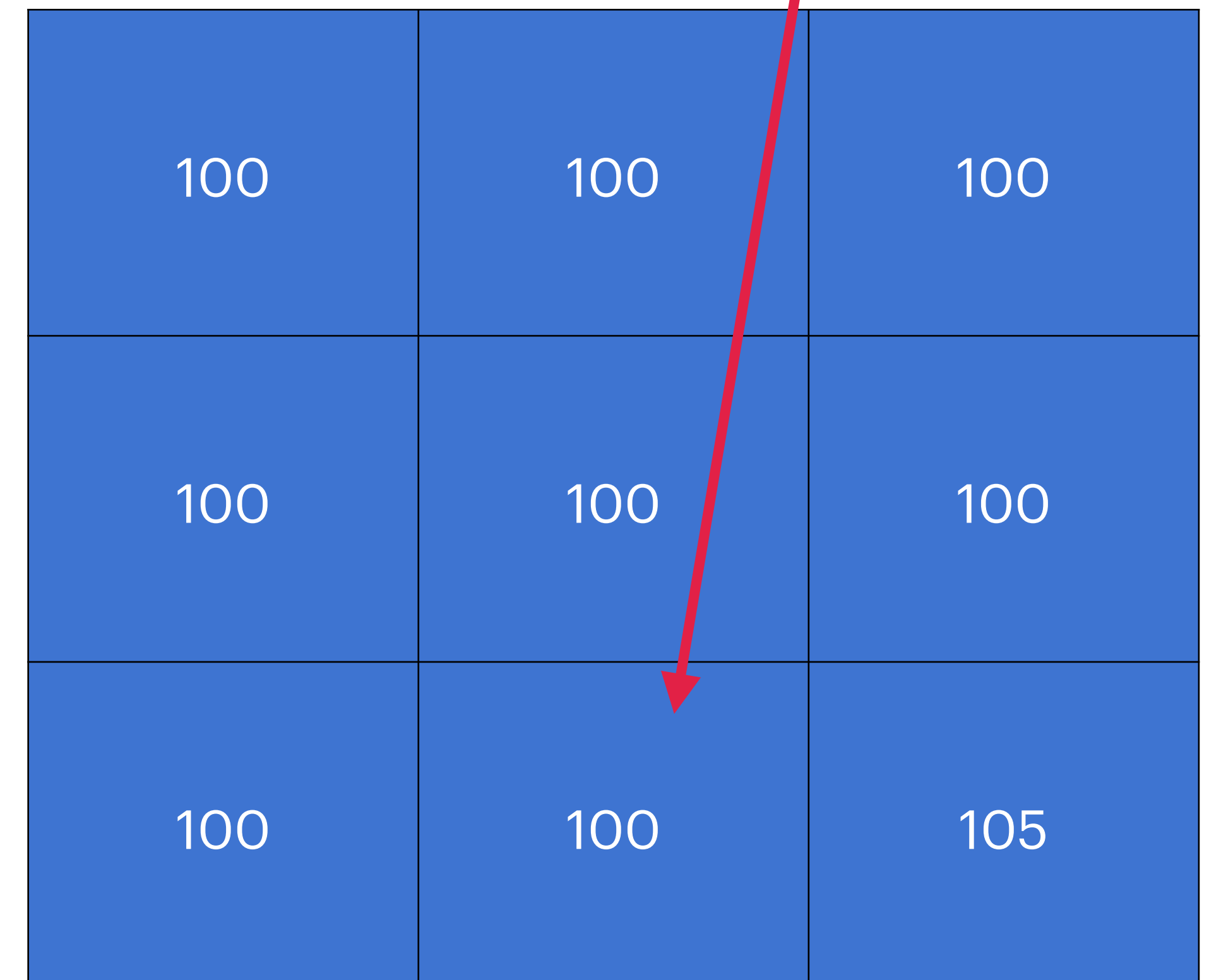
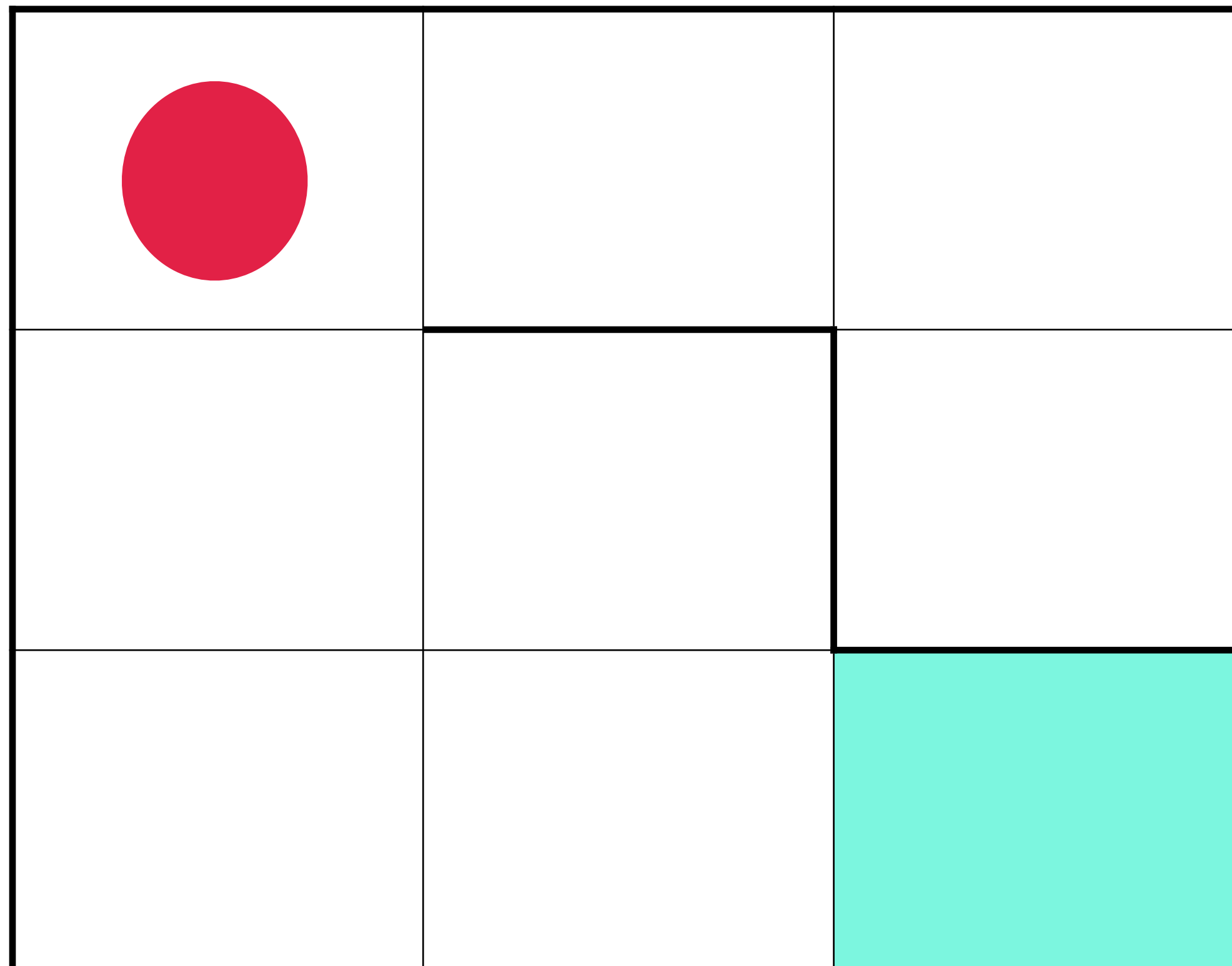
$$v^*(s) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v^*(s')]$$



Exemplo visual

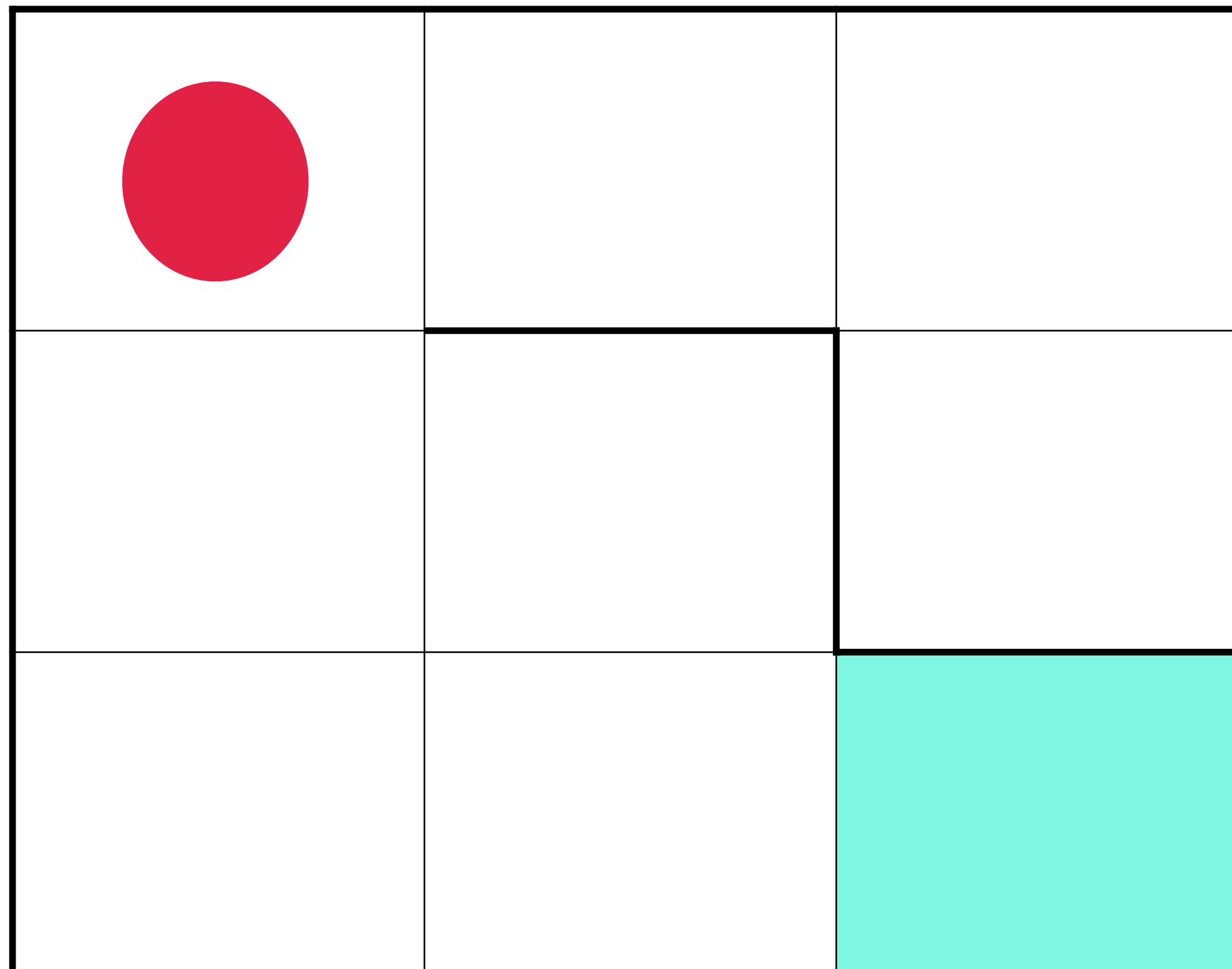
- Para valores $v(s)$, $t = 1$

$$v^*(s) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v^*(s')]$$

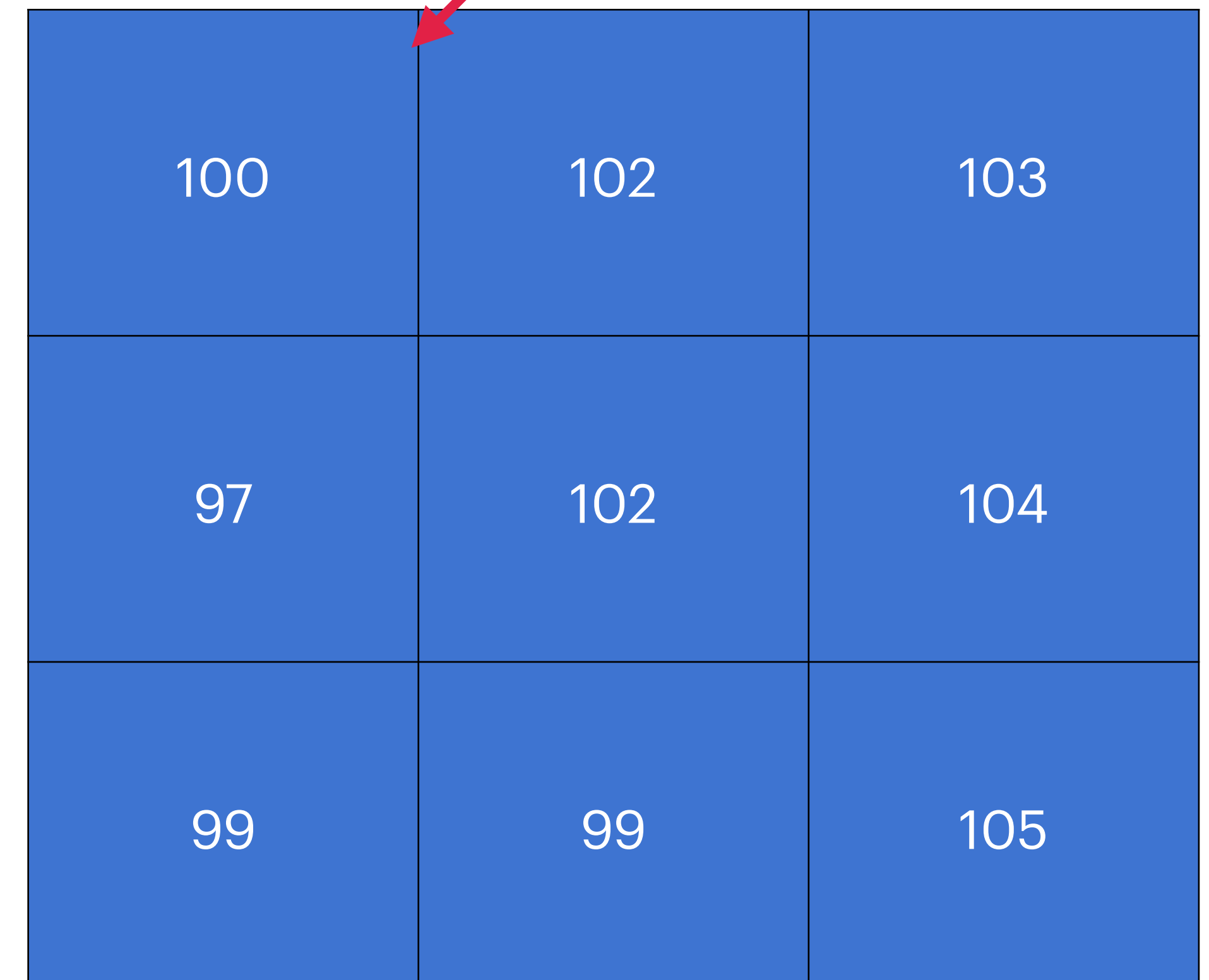


Exemplo visual

- Para valores $v(s)$, $t = 1$

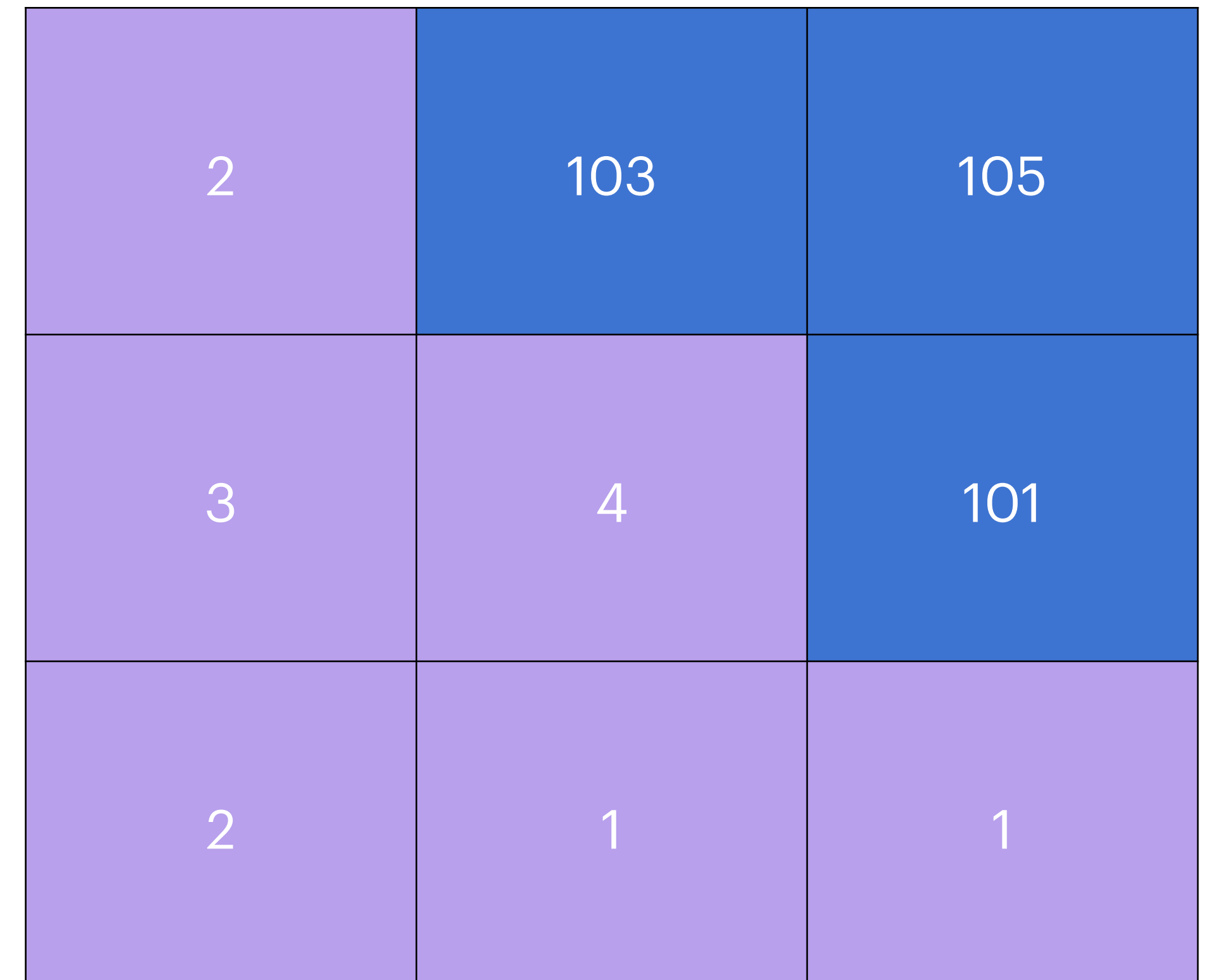
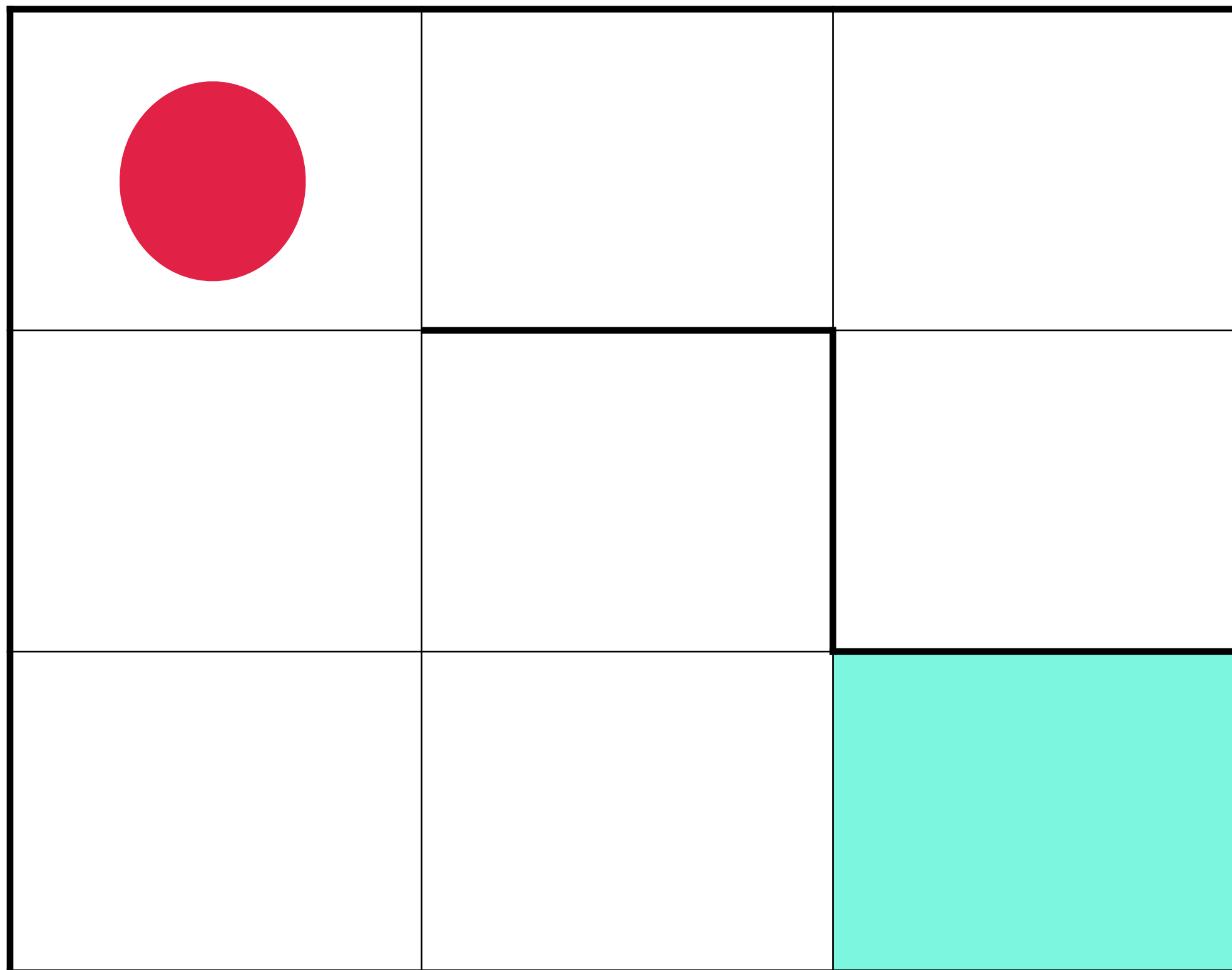


$$v^*(s) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v^*(s')]$$



Exemplo visual

- Para valores $v(s), t = n$



Resolvendo um MDP

Abra o Notebook MDP.ipynb