UFRN

metrópole DIGITAL

Aprendizado por Reforço

Processos de Decisão de Markov
(parte 2)

Recapitulação da aula passada...

- Trajetória Estado-Ação-Recompensa: $S_t \xrightarrow{A_t} S_{t+1}, R_{t+1} \xrightarrow{A_{t+1}} S_{t+2}, R_{t+2} \xrightarrow{A_{t+2}} S_{t+3}, R_{t+3} \cdots$
- Objetivo: maximizar recompensas acumulada a longo prazo <u>e não recompensas imediatas.</u>
- Retorno descontado

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=t+1}^{T} \gamma^{k-(t+1)} R_k$$

  - Taxa de desconto (pondera recompensas futuras): $0 \leq \gamma \leq 1$ ($\gamma = 1$ ou $T = \infty$, <u>mas não ambos</u>)
  - $S_t, S_{t+1} \in \mathcal{S}$: variáveis aleatórias
  - $A_t \in \mathcal{A}(S_t)$: variável aleatória
  - $R_{t+1} \in \mathcal{R}(S_t, A_t)$: variável aleatória
  - G_t: variável aleatória

- Função de valor de estado (ou valor de estado de s) para a política π

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

Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Ferramenta fundamental para projetar e analisar algoritmos de aprendizado por reforço.
  - Sistema de equações lineares que descreve os valores de todos os estados
  - Retorno descontado (computação recursiva):

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$G_t = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)$$

$$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots$$

$$\boxed{G_t = R_{t+1} + \gamma G_{t+1}}$$

Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Valor de estado

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$
$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$$
$$v_\pi(s) = \underbrace{\mathbb{E}_\pi[R_{t+1}|S_t = s]} + \gamma \underbrace{\mathbb{E}_\pi[G_{t+1}|S_t = s]}$$

$$\mathbb{E}_\pi[R_{t+1}|S_t = s] = \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_\pi[R_{t+1}|S_t = s, A_t = a]$$
$$\mathbb{E}_\pi[R_{t+1}|S_t = s] = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{r\in\mathcal{R}} p(r|s,a)r$$

Valor esperado das recompensas imediatas

$$\mathbb{E}_\pi[G_{t+1}|S_t = s] = \sum_{s'\in\mathcal{S}} \mathbb{E}_\pi[G_{t+1}|S_t = s, S_{t+1} = s']p(s'|s)$$
$$\mathbb{E}_\pi[G_{t+1}|S_t = s] = \sum_{s'\in\mathcal{S}} \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']p(s'|s)$$
$$\mathbb{E}_\pi[G_{t+1}|S_t = s] = \sum_{s'\in\mathcal{S}} v_\pi(s')p(s'|s)$$
$$\mathbb{E}_\pi[G_{t+1}|S_t = s] = \sum_{s'\in\mathcal{S}} v_\pi(s') \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s)$$

Valor esperado das recompensas futuras

Valores de Estado e Equação de Bellman

- Equação de Bellman

$$
\begin{aligned}
\boxed{v_\pi(s)} &= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s], \\
&= \underbrace{\sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s,a) r}_{\text{mean of immediate rewards}} + \underbrace{\gamma \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s'|s,a) v_\pi(s')}_{\text{mean of future rewards}} \\
&= \sum_{a \in \mathcal{A}} \boxed{\pi(a|s)} \left[\sum_{r \in \mathcal{R}} \boxed{p(r|s,a)} r + \gamma \sum_{s' \in \mathcal{S}} \boxed{p(s'|s,a)} \boxed{v_\pi(s')} \right], \quad \text{for all } s \in \mathcal{S}.
\end{aligned}
$$

Modelo do ambiente

Valores desconhecidos que queremos calcular!

Resolver o sistema de equações e encontrar os valores de estado significa avaliar a política π

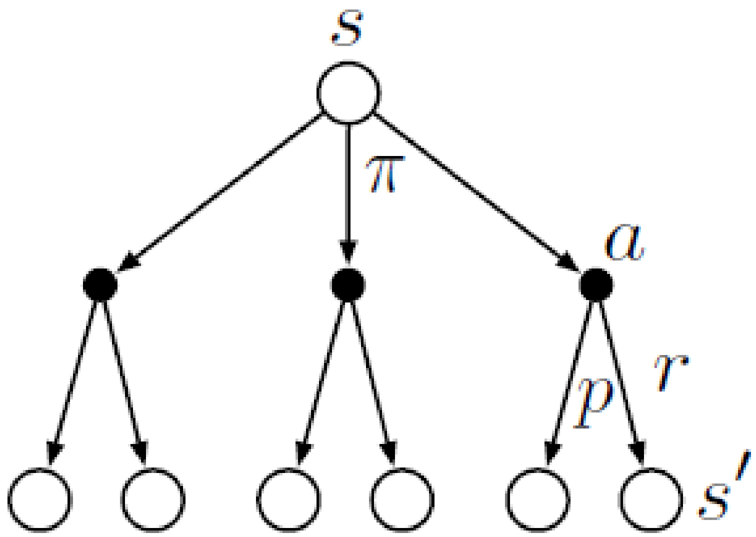
Valores de Estado e Equação de Bellman

- Equação de Bellman

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

$$p(s'|s, a) = \sum_{r \in \mathcal{R}} p(s', r|s, a)$$

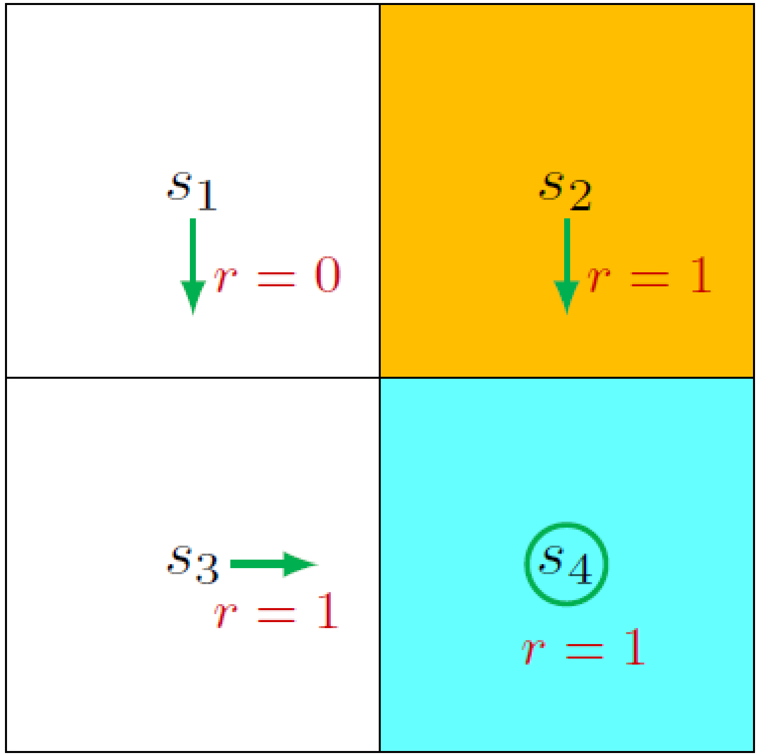
$$p(r|s, a) = \sum_{s' \in \mathcal{S}} p(s', r|s, a)$$

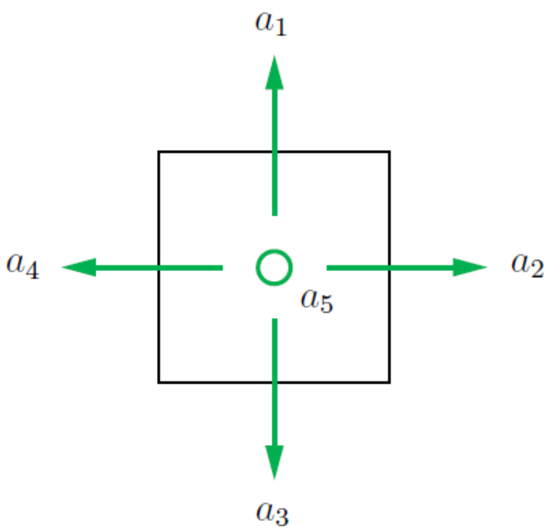


Backup diagram for v_π

Valores de Estado e Equação de Bellman

- Equação de Bellman





Política e ambiente determinísticos

$$\pi(a = a_3|s_1) = 1$$

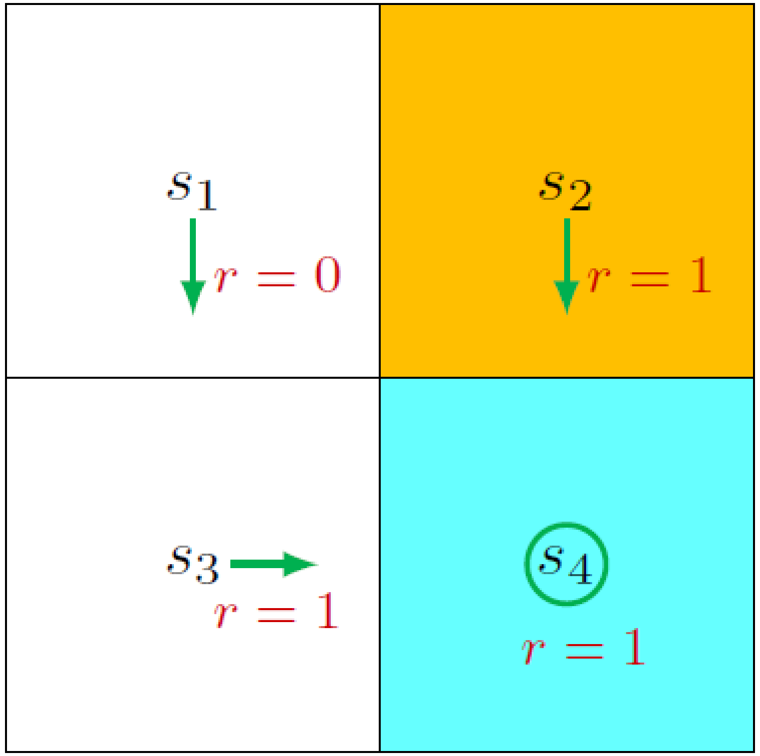
$$\pi(a \neq a_3|s_1) = 0$$

$$p(s' = s_3|s_1, a_3) = 1$$

$$p(s' \neq s_3|s_1, a_3) = 0$$

Valores de Estado e Equação de Bellman

- Equação de Bellman



$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

$$v_\pi(s_1) = 0 + \gamma v_\pi(s_3)$$

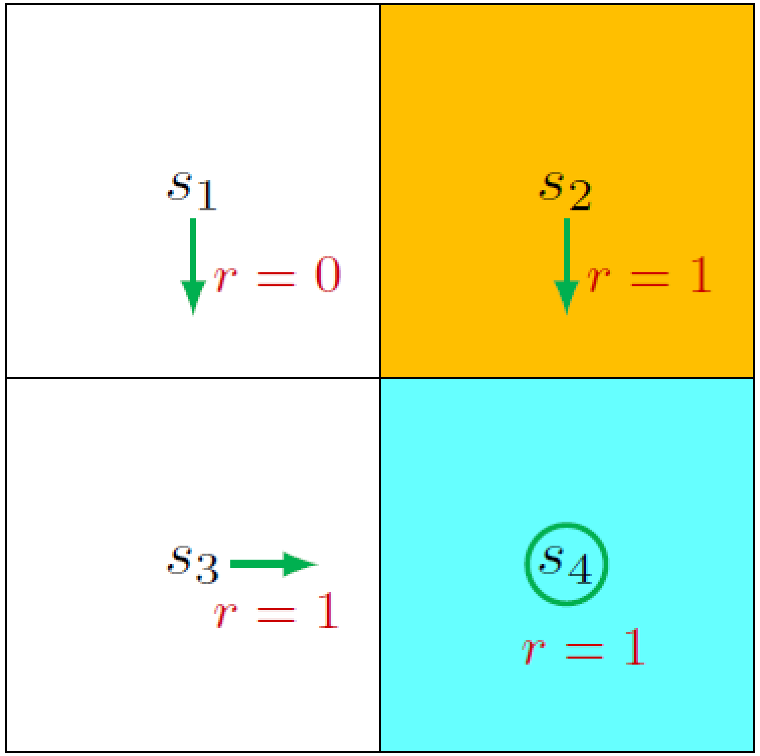
$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

Valores de Estado e Equação de Bellman

- Equação de Bellman



$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

$$v_\pi(s_1) = 0 + \gamma v_\pi(s_3)$$

$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_1) = \frac{\gamma}{1-\gamma}$$

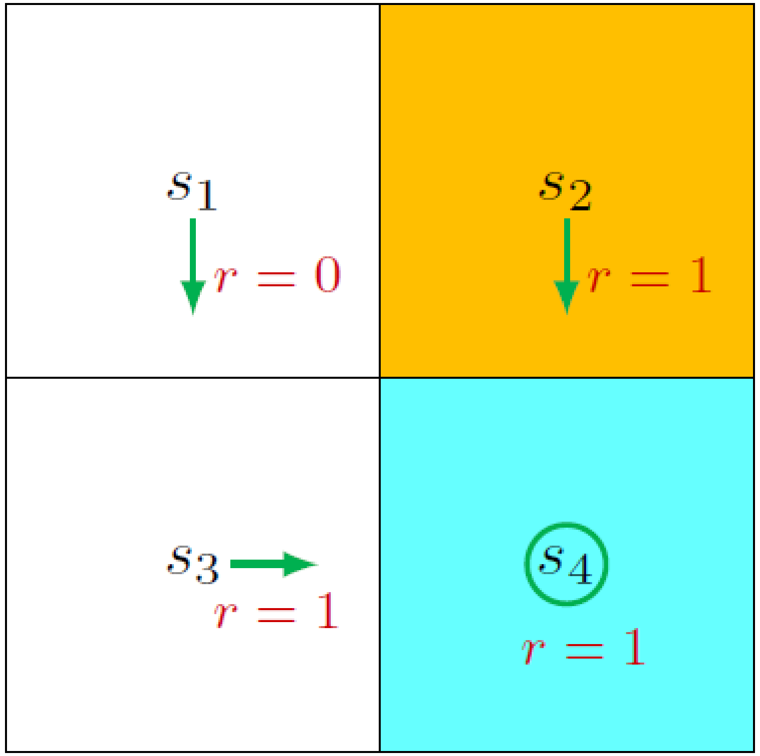
$$v_\pi(s_2) = \frac{1}{1-\gamma}$$

$$v_\pi(s_3) = \frac{1}{1-\gamma}$$

$$v_\pi(s_4) = \frac{1}{1-\gamma}$$

Valores de Estado e Equação de Bellman

- Equação de Bellman



$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

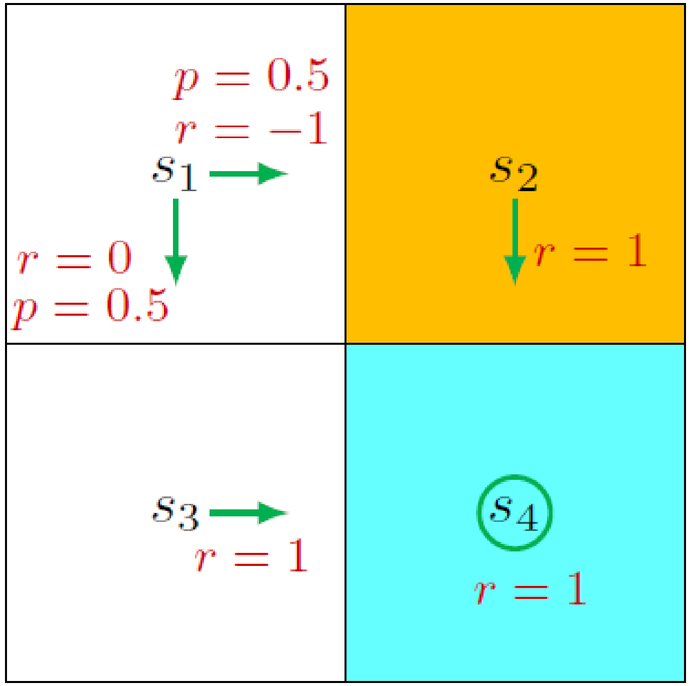
$$v_\pi(s_1) = 0 + \gamma v_\pi(s_3)$$
$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

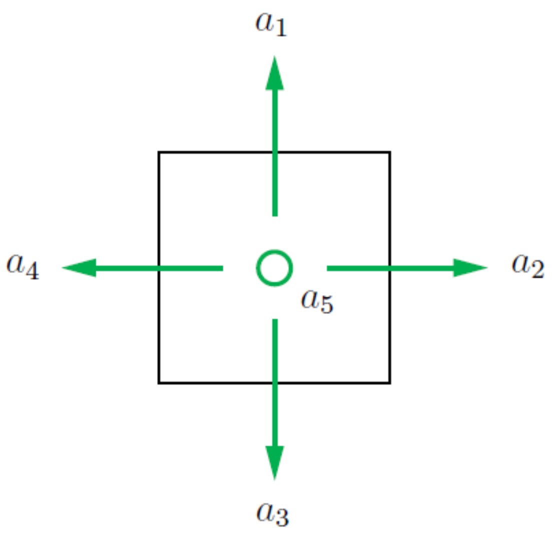
$$v_\pi(s_1) = \frac{\gamma}{1-\gamma}$$
$$v_\pi(s_2) = \frac{1}{1-\gamma}$$
$$v_\pi(s_3) = \frac{1}{1-\gamma}$$
$$v_\pi(s_4) = \frac{1}{1-\gamma}$$

$$v_\pi(s_1) = \frac{0.9}{1-0.9} = 9$$
$$v_\pi(s_2) = \frac{1}{1-0.9} = 10$$
$$v_\pi(s_3) = \frac{1}{1-0.9} = 10$$
$$v_\pi(s_4) = \frac{1}{1-0.9} = 10$$

Valores de Estado e Equação de Bellman

- Equação de Bellman





Política estocástica

$\pi(a = a_2|s_1) = 0.5$

$\pi(a = a_3|s_1) = 0.5$

Ambiente determinístico

$p(s' = s_3|s_1, a_3) = 1$

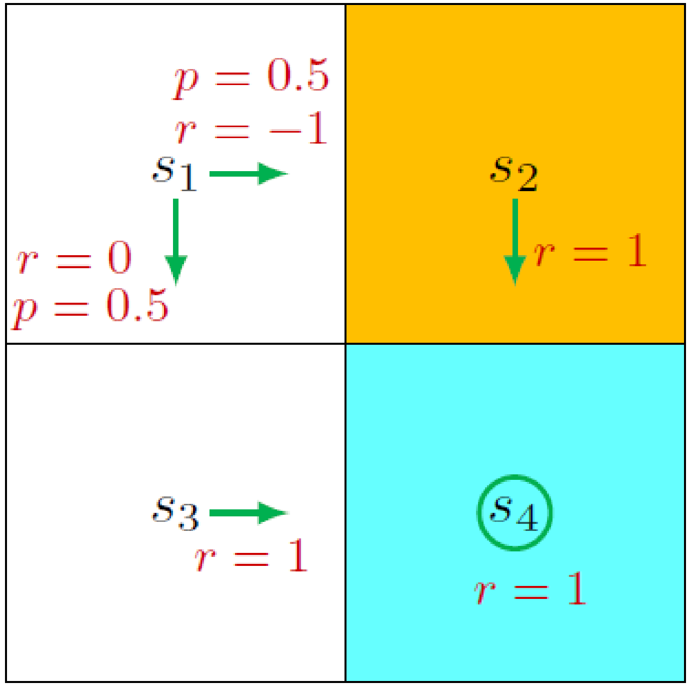
$p(s' = s_2|s_1, a_2) = 1$

$p(r = 0|s_1, a_3) = 1$

$p(r = -1|s_1, a_2) = 1$

Valores de Estado e Equação de Bellman

- Equação de Bellman



$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

$$v_\pi(s_1) = 0.5[0 + \gamma v_\pi(s_3)] + 0.5[-1 + \gamma v_\pi(s_2)]$$

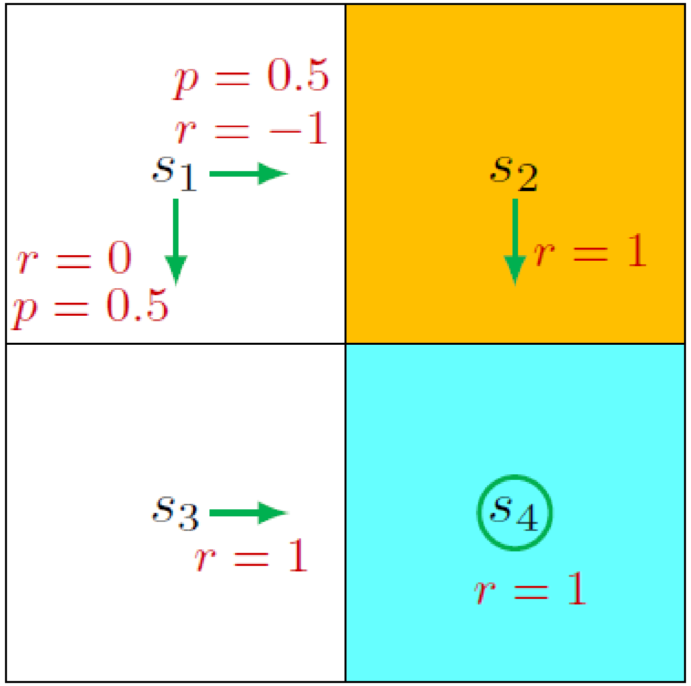
$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

Valores de Estado e Equação de Bellman

- Equação de Bellman



$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

$$v_\pi(s_1) = 0.5[0 + \gamma v_\pi(s_3)] + 0.5[-1 + \gamma v_\pi(s_2)]$$

$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

$$v_\pi(s_1) = -0.5 + \frac{\gamma}{1-\gamma}$$

$$v_\pi(s_2) = \frac{1}{1-\gamma}$$

$$v_\pi(s_3) = \frac{1}{1-\gamma}$$

$$v_\pi(s_4) = \frac{1}{1-\gamma}$$

$$v_\pi(s_1) = -0.5 + 9 = 8.5$$

$$v_\pi(s_2) = 10$$

$$v_\pi(s_3) = 10$$

$$v_\pi(s_4) = 10$$

Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Forma escalar

$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

  - Forma matricial

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s' \in \mathcal{S}} p_\pi(s'|s) v_\pi(s')$$

    - Onde,

$$r_\pi(s) \triangleq \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s,a)r$$

$$p_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a)$$

Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Forma matricial (continuação)
    - Considerando os estados indexados por $i = 1, \dots, |\mathcal{S}|$

$$v_\pi(s_i) = r_\pi(s_i) + \gamma \sum_{s_j \in \mathcal{S}} p_\pi(s_j|s_i) v_\pi(s_j)$$

    - Seja

$$v_\pi = [v_\pi(s_1), \cdots, v_\pi(s_n)]^\top \in \mathbb{R}^n$$

$$r_\pi = [r_\pi(s_1), \cdots, r_\pi(s_n)]^\top \in \mathbb{R}^n$$

$$P_\pi \in \mathbb{R}^{n \times n}, \qquad [P_\pi]_{ij} = p_\pi(s_j|s_i)$$

$$\boxed{v_\pi = r_\pi + \gamma P_\pi v_\pi}$$

conhecido

desconhecido

Propriedades da matrix P

Elementos $\mathrm{p}_{i,j} \geq 0$

$$\sum_i p_{i,j} = 1$$

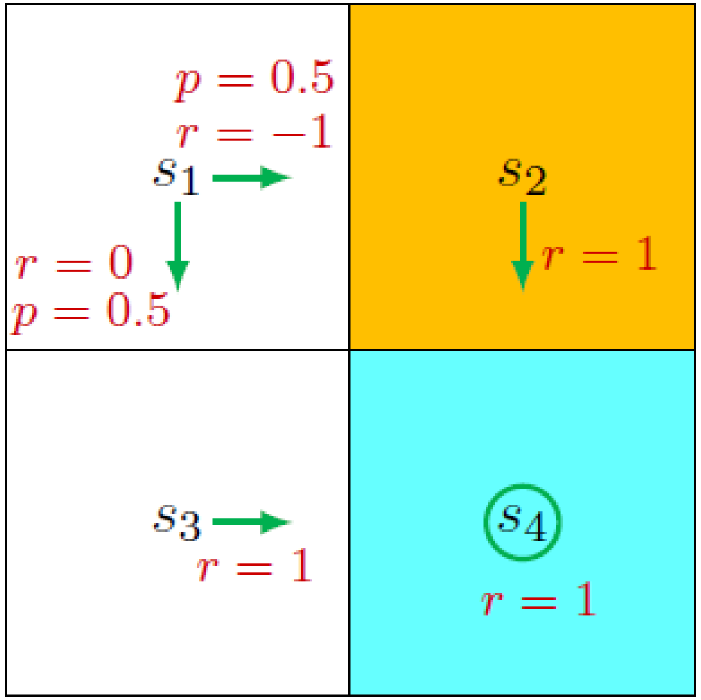
Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Forma matricial (continuação)

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

$$r_\pi(s) \triangleq \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s,a) r$$

$$p_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a)$$



$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}$$

$$\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0.5(0) + 0.5(-1) \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}$$

Valores de Estado e Equação de Bellman

- Equação de Bellman
  - Avaliação de uma política: Calcular os valores de estado dado a política
  - Cálculo dos valores de estado
    - Solução analítica

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

    - Solução iterativa

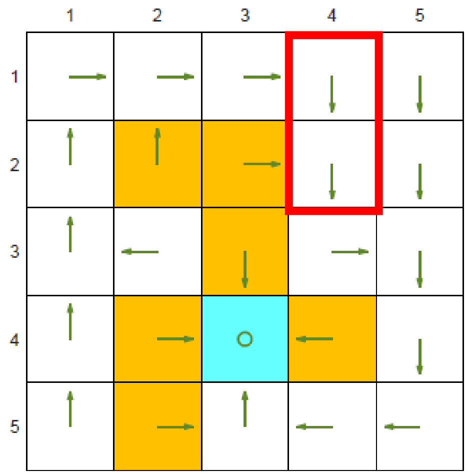
$$v_{k+1} = r_\pi + \gamma P_\pi v_k, \qquad k = 0,1,2, \dots$$

Esse algoritmo gera uma sequência de valores $\{v_0, v_1, \dots, \}$ que converge para v_π

$$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi, \qquad k \to \infty$$

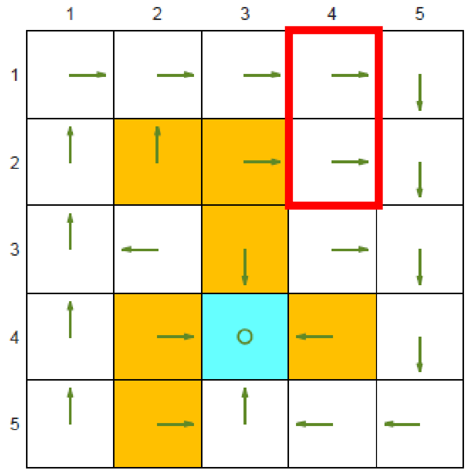
Valores de Estado e Equação de Bellman

- Equação de Bellman



	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

Exemplos de boas políticas eseus valores de estado

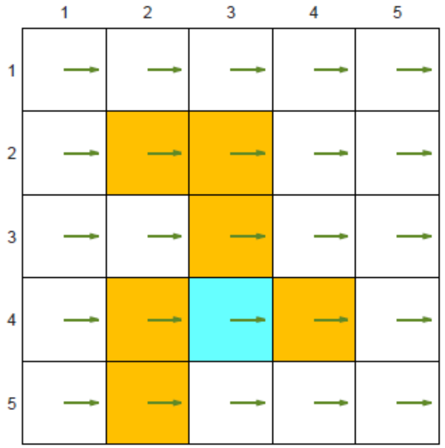


	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

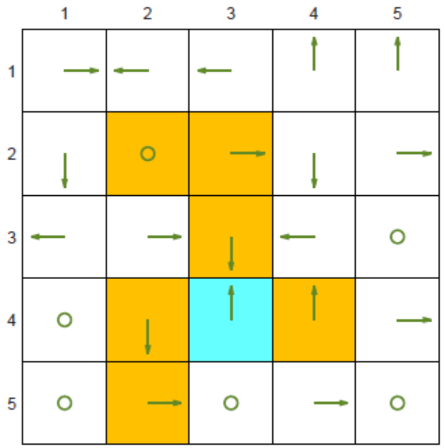
Políticas distintas podem ter os mesmos valores de estado

Valores de Estado e Equação de Bellman

- Equação de Bellman



	1	2	3	4	5
1	-6.6	-7.3	-8.1	-9.0	-10.0
2	-8.5	-8.3	-8.1	-9.0	-10.0
3	-7.5	-8.3	-8.1	-9.0	-10.0
4	-7.5	-7.2	-9.1	-9.0	-10.0
5	-7.6	-7.3	-8.1	-9.0	-10.0



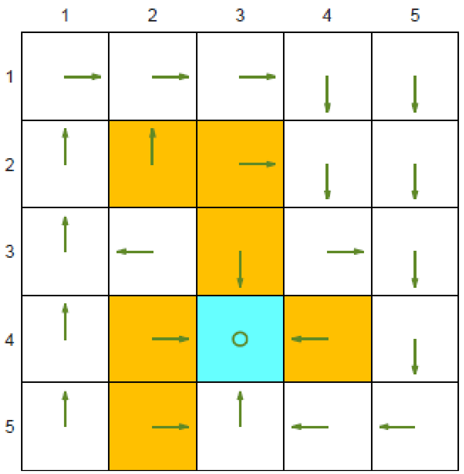
	1	2	3	4	5
1	0.0	0.0	0.0	-10.0	-10.0
2	-9.0	-10.0	-0.4	-0.5	-10.0
3	-10.0	-0.5	0.5	-0.5	0.0
4	0.0	-1.0	-0.5	-0.5	-10.0
5	0.0	0.0	0.0	0.0	0.0

Exemplos de políticas ruins e seus valores de estado

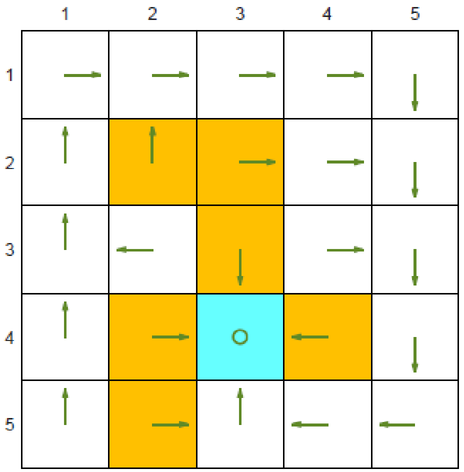
Valores de Estado e Equação de Bellman

- Equação de Bellman

Políticas boas

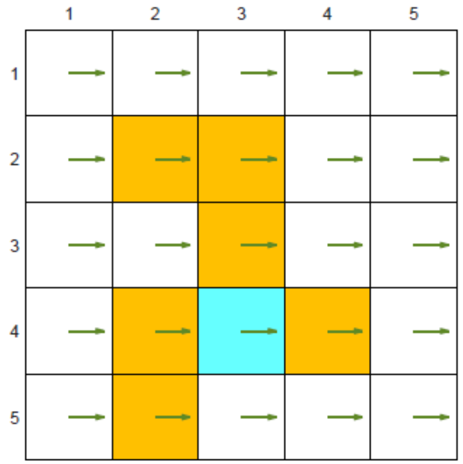


	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

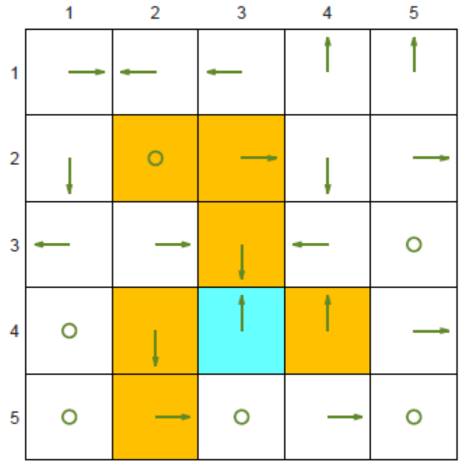


	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

Políticas ruins



	1	2	3	4	5
1	-6.6	-7.3	-8.1	-9.0	-10.0
2	-8.5	-8.3	-8.1	-9.0	-10.0
3	-7.5	-8.3	-8.1	-9.0	-10.0
4	-7.5	-7.2	-9.1	-9.0	-10.0
5	-7.6	-7.3	-8.1	-9.0	-10.0



	1	2	3	4	5
1	0.0	0.0	0.0	-10.0	-10.0
2	-9.0	-10.0	-0.4	-0.5	-10.0
3	-10.0	-0.5	0.5	-0.5	0.0
4	0.0	-1.0	-0.5	-0.5	-10.0
5	0.0	0.0	0.0	0.0	0.0

Valores de Estado e Equação de Bellman

- Valores de Ação
  - Indicam o "valor" de executar uma ação em um estado

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

Valores de Estado e Equação de Bellman

- Relação entre valor de estado e valor de ação

1. Obtendo o valor de estado a partir do valor de ação

$$\underbrace{\mathbb{E}[G_t|S_t = s]}_{v_\pi(s)} = \sum_{a\in\mathcal{A}} \underbrace{\mathbb{E}[G_t|S_t = s, A_t = a]}_{q_\pi(s,a)}\pi(a|s) \longrightarrow v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\underbrace{q_\pi(s,a)}$$

- O valor de estado é o valor esperado dos valores de ação possíveis naquele estado, segundo a política π.

2. Obtendo o valor de ação a partir do valor de estado

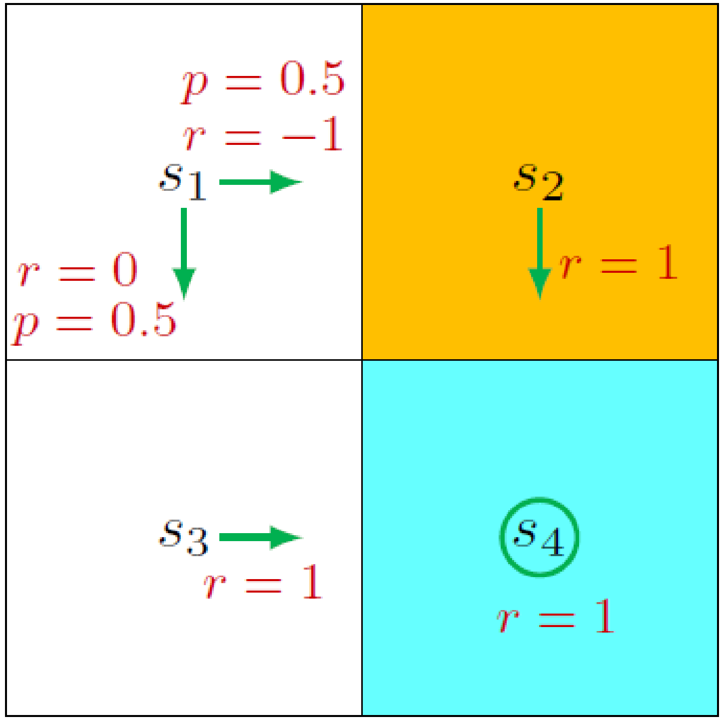
$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\underbrace{\left[\sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\, v_\pi(s')\right]} \longrightarrow q_\pi(s,a) = \sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\, v_\pi(s')$$

- O valor de ação é a soma do valor esperado de recompensa imediata ao executar a ação a no estado s com o valor esperado de recompensas futuras.

Valores de Estado e Equação de Bellman

- Valores de Ação

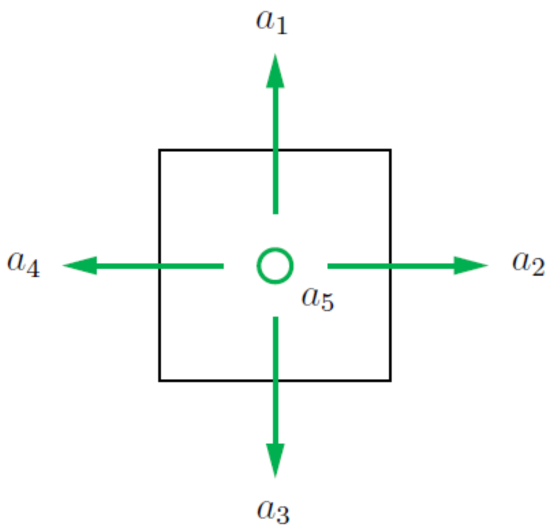
$$q_\pi(s,a) = \sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\, v_\pi(s')$$



$$q_\pi(s_1,a_2) = -1 + \gamma v_\pi(s_2)$$

$$q_\pi(s_1,a_3) = 0 + \gamma v_\pi(s_3)$$

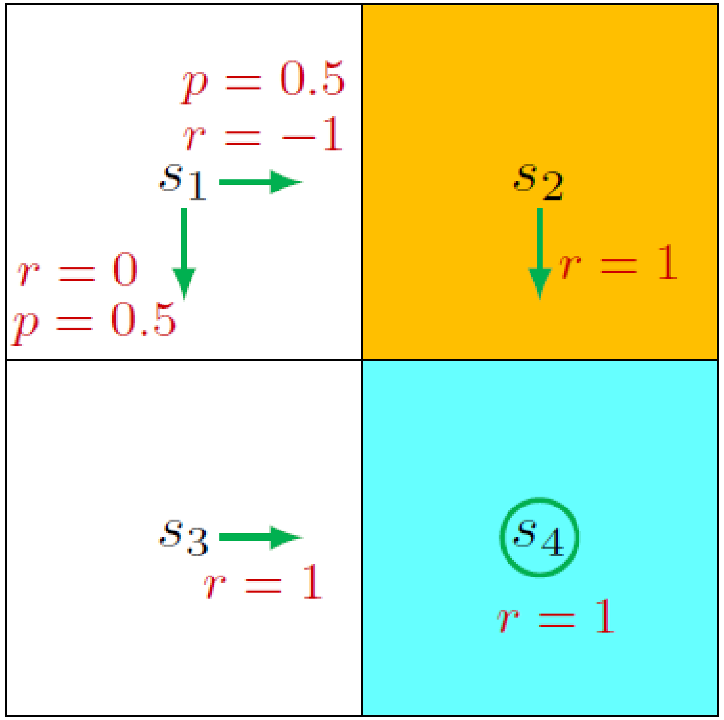
Mas e as outras ações?



Valores de Estado e Equação de Bellman

- Valores de Ação

$$q_\pi(s,a) = \sum_{r \in \mathcal{R}} p(r|s,a) r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)\, v_\pi(s')$$



$$q_\pi(s_1, a_2) = -1 + \gamma v_\pi(s_2)$$

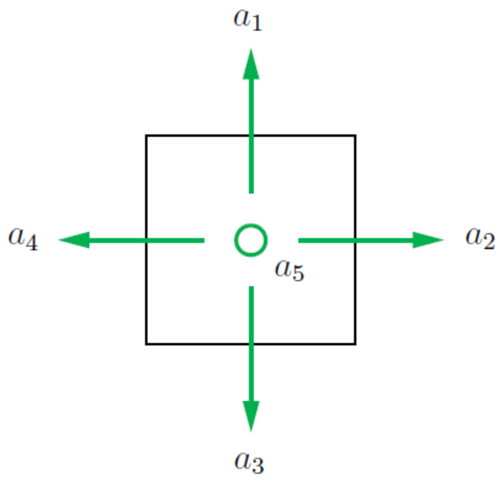
$$q_\pi(s_1, a_3) = 0 + \gamma v_\pi(s_3)$$

Mas e as outras ações?

$$q_\pi(s_1, a_1) = -1 + \gamma v_\pi(s_1)$$

$$q_\pi(s_1, a_4) = -1 + \gamma v_\pi(s_1)$$

$$q_\pi(s_1, a_5) = 0 + \gamma v_\pi(s_1)$$



Mesmo que a política não selecione uma ação (a_1, a_4, a_5), essas ações ainda têm valores de ação.

Valores de Estado e Equação de Bellman

- Valores de Ação
  - Mesmo que a política atual não selecione certas ações, elas podem ser melhores do que as ações escolhidas pela política.
  - O objetivo do aprendizado por reforço é encontrar políticas ótimas, por isso precisamos continuar explorando todas as ações para determinar a melhor ação para cada estado.
  - Cálculo do valor de estado a partir dos valores de ação
    - Uma vez calculados os valores de ação, é possível obter o valor de estado via

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$$

$$v_\pi(s_1) = 0.5 q_\pi(s_1, a_2) + 0.5 q_\pi(s_1, a_3)$$

$$v_\pi(s_1) = 0.5[0 + \gamma v_\pi(s_3)] + 0.5[-1 + \gamma v_\pi(s_2)]$$

Valores de Estado e Equação de Bellman

- Valores de Ação
  - Equação de Bellman em termos de valores de ação

$$q_\pi(s,a) = \sum_{r \in \mathcal{R}} p(r|s,a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \sum_{a' \in \mathcal{A}(s')} \pi(a'|s') q_\pi(s',a')$$

Referências

- Shiyu Zhao. Mathematical Foundations of Reinforcement Learning. Springer Singapore, 2025. [capítulo 2]
  - disponível em: https://github.com/MathFoundationRL/Book-Mathematical-Foundation-of-Reinforcement-Learning

- Richard S. Sutton e Andrew G. Barto. An Introduction Reinforcement Learning, Bradford Book, 2018. [capítulo 3]
  - disponível em: http://incompleteideas.net/book/the-book-2nd.html

Slides construídos com base nos livros supracitados, os quais estão disponibilizados publicamente pelos seus respectivos autores.