Aprendizado por Reforço

Programação dinâmica (parte 1)

Recapitulação das aulas passadas

- Processos de decisão de Markov
- Valores de estado
- Valores de ação
- Equação de Bellman
- Equação de otimalidade de Bellman
- Valores de estado ótimos
- Política ótima

Algoritmos

- Alguns algoritmos de programação dinâmica para encontrar políticas ótimas
  - Iteração de valor
  - Iteração de política
  - Iteração de política truncada
- Os algoritmos acima exigem **modelo do sistema** (isto é, assumem o conhecimento do ambiente)
- Desafio: não é prático para problemas muito grandes

Iteração de valor

- Algoritmo sugerido pelo teorema do ponto fixo

$$v_{k+1} = f(v_k) = \max_{\pi \in \Pi}(r_\pi + \gamma P_\pi v_k), \qquad k = 0, 1, 2, \ldots$$

- É garantido que $v_k \to v^*$ e $\pi_k \to \pi^*$ quando $k \to \infty$.

- O algoritmo tem 2 passos em cada iteração
  1. Atualização de política
  2. Atualização de valor

Iteração de valor (forma matricial)

- **Passo 1**: Atualização de Política

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}(r_\pi + \gamma P_\pi v_k)$$

v_k: obtido da iteração anterior

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

$$[P_\pi]_{s,s'} = p(s'|s) \triangleq \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a)$$

- **Passo 2**: Atualização de Valor

$$v_{k+1} = r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} v_k$$

v_{k+1}: vai ser utilizado na próxima iteração

Iteração de valor (forma escalar)

- **Passo 1**: Atualização de política

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}} \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right], \qquad s \in \mathcal{S}$$

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}} \sum_a \pi(a|s) q_k(s,a) \ , \quad s \in \mathcal{S}$$

- A política ótima (**gulosa**) que resolve esse problema de otimização é:

$$\boxed{\pi(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases}}, \qquad onde \qquad a_k^*(s) = \underset{a}{\operatorname{argmax}}\, q(s,a)$$

Iteração de valor (forma escalar)

- **Passo 2**: Atualização de valor

$$v_{k+1}(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right], \qquad s \in \mathcal{S}$$

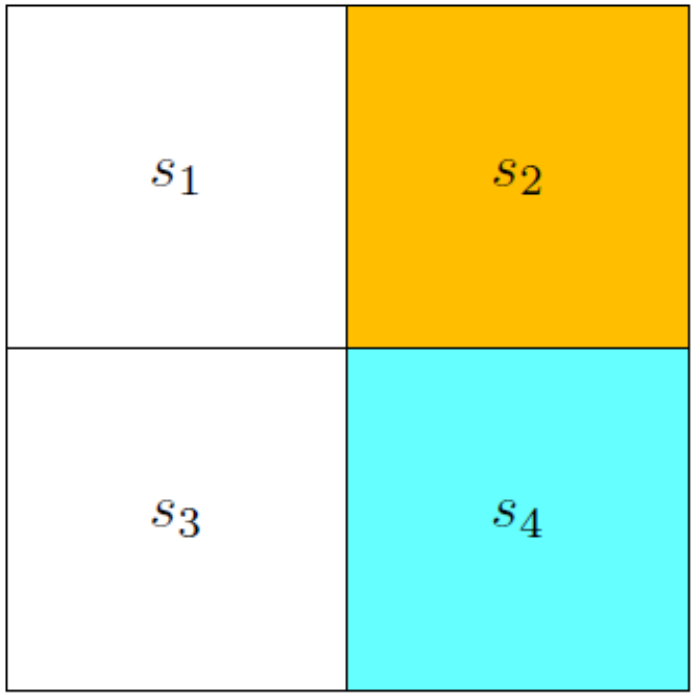
$$v_{k+1}(s) = \sum_a \pi(a|s) q_k(s,a) , \qquad s \in \mathcal{S}$$

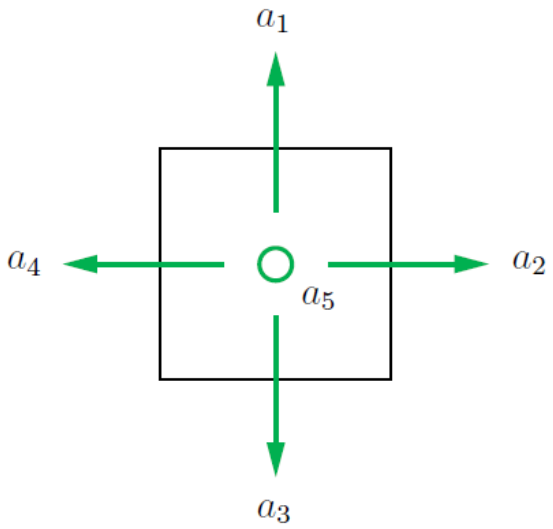
- De acordo com o passo 1:

$$\boxed{v_{k+1}(s) = \max_a q_k(s,a)}, \qquad s \in \mathcal{S}$$

Iteração de valor

- Exemplo





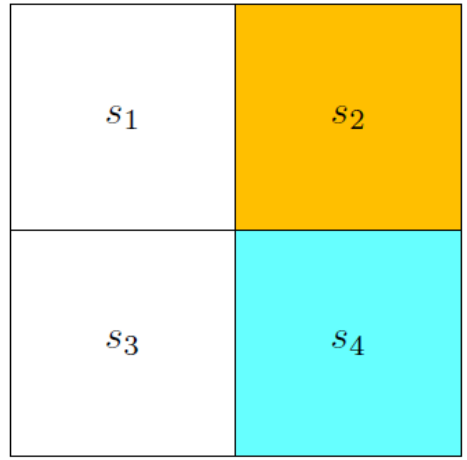
$$r_{boundary} = r_{forbidden} = -1\,, r_{target} = 1,\ \gamma = 0.9$$

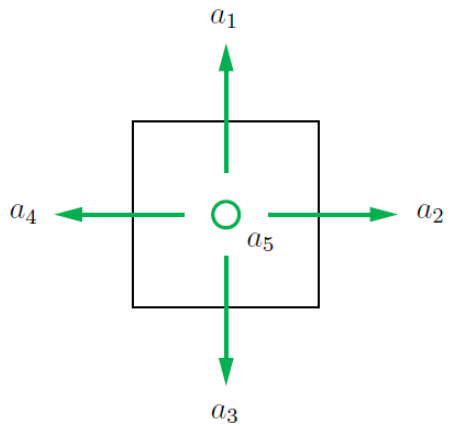
Iteração de valor (forma escalar)

$k = 0$ e $v_0(s_1) = v_0(s_2) = v_0(s_3) = v_0(s_4) = 0$

$$q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s')$$

q-table	a_1	a_2	a_3	a_4	a_5
s_1					
s_2					
s_3					
s_4					

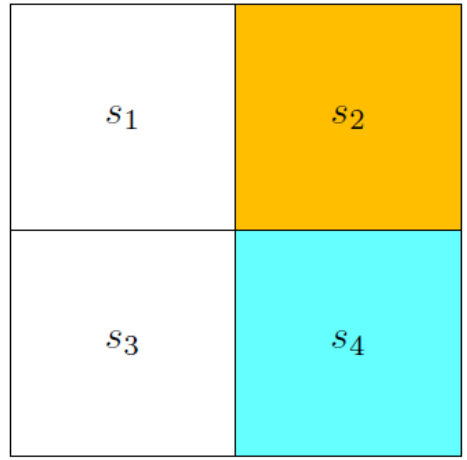




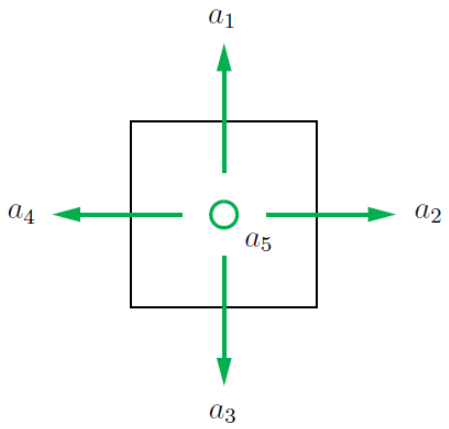
Iteração de valor (forma escalar)

$$k = 0 \text{ e } v_0(s_1) = v_0(s_2) = v_0(s_3) = v_0(s_4) = 0$$

$$q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s')$$



q-table	a_1	a_2	a_3	a_4	a_5
s_1	$-1+\gamma v(s_1)$	$-1+\gamma v(s_2)$	$0+\gamma v(s_3)$	$-1+\gamma v(s_1)$	$0+\gamma v(s_1)$
s_2	$-1+\gamma v(s_2)$	$-1+\gamma v(s_2)$	$1+\gamma v(s_4)$	$0+\gamma v(s_1)$	$-1+\gamma v(s_2)$
s_3	$0+\gamma v(s_1)$	$1+\gamma v(s_4)$	$-1+\gamma v(s_3)$	$-1+\gamma v(s_3)$	$0+\gamma v(s_3)$
s_4	$-1+\gamma v(s_2)$	$-1+\gamma v(s_4)$	$-1+\gamma v(s_4)$	$0+\gamma v(s_3)$	$1+\gamma v(s_4)$

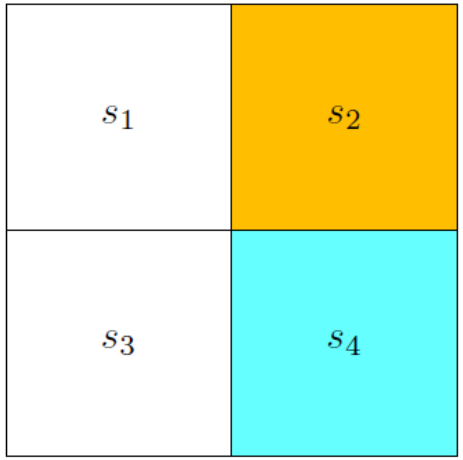


q-table	a_1	a_2	a_3	a_4	a_5
s_1	-1	-1	0	-1	0
s_2	-1	-1	1	0	-1
s_3	0	1	-1	-1	0
s_4	-1	-1	-1	0	1

Iteração de valor

Passo 1: Atualização de política

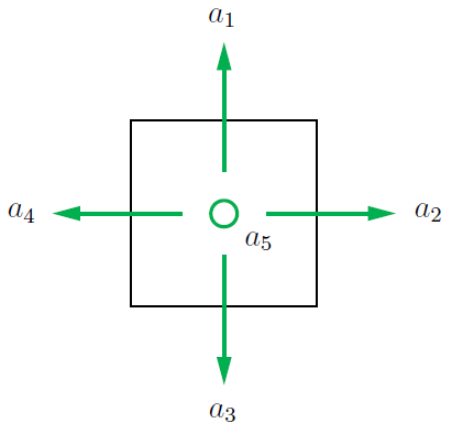
q-table	a_1	a_2	a_3	a_4	a_5
s_1	-1	-1	0	-1	0
s_2	-1	-1	1	0	-1
s_3	0	1	-1	-1	0
s_4	-1	-1	-1	0	1



$$\pi_1(a_5|s_1) = 1$$
$$\pi_1(a_3|s_2) = 1$$
$$\pi_1(a_2|s_3) = 1$$
$$\pi_1(a_5|s_4) = 1$$

$$\pi(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases}$$

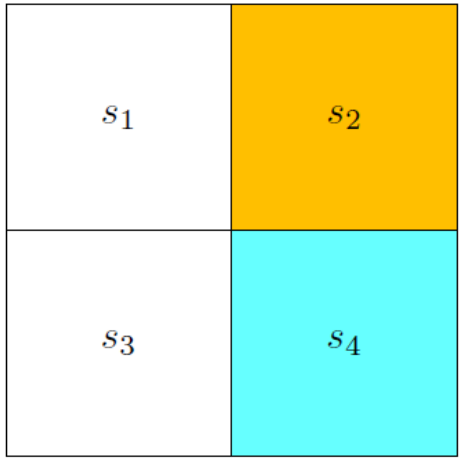
$$onde \quad a_k^*(s) = \underset{a}{\operatorname{argmax}}\, q(s,a)$$



Iteração de valor

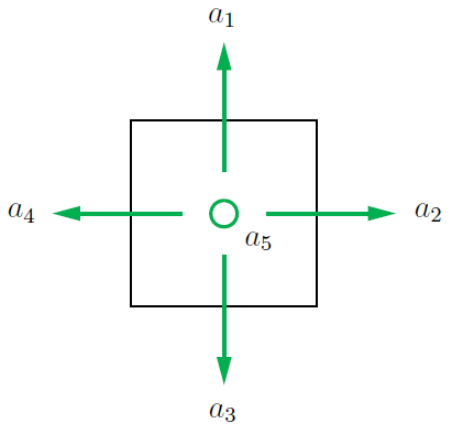
Passo 2: Atualização de valor

q-table	a_1	a_2	a_3	a_4	a_5
s_1	-1	-1	0	-1	0
s_2	-1	-1	1	0	-1
s_3	0	1	-1	-1	0
s_4	-1	-1	-1	0	1



$$v_1(s_1) = 0$$
$$v_1(s_2) = 1$$
$$v_1(s_3) = 1$$
$$v_1(s_4) = 1$$

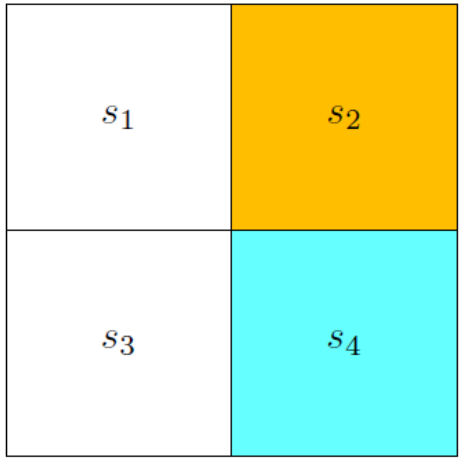
$$v_{k+1}(s) = \max_a q_k(s, a)$$



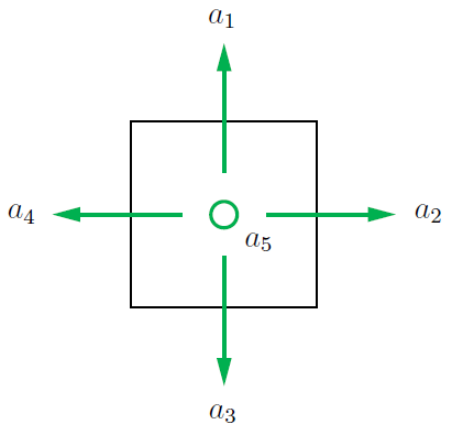
Iteração de valor

$$k = 1 \text{ e } v_1(s_1) = 0, v_1(s_2) = 1, v_1(s_3) = 1, v_1(s_4) = 1$$

$$q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s')$$



q-table	a_1	a_2	a_3	a_4	a_5
s_1	$-1+\gamma v(s_1)$	$-1+\gamma v(s_2)$	$0+\gamma v(s_3)$	$-1+\gamma v(s_1)$	$0+\gamma v(s_1)$
s_2	$-1+\gamma v(s_2)$	$-1+\gamma v(s_2)$	$1+\gamma v(s_4)$	$0+\gamma v(s_1)$	$-1+\gamma v(s_2)$
s_3	$0+\gamma v(s_1)$	$1+\gamma v(s_4)$	$-1+\gamma v(s_3)$	$-1+\gamma v(s_3)$	$0+\gamma v(s_3)$
s_4	$-1+\gamma v(s_2)$	$-1+\gamma v(s_4)$	$-1+\gamma v(s_4)$	$0+\gamma v(s_3)$	$1+\gamma v(s_4)$

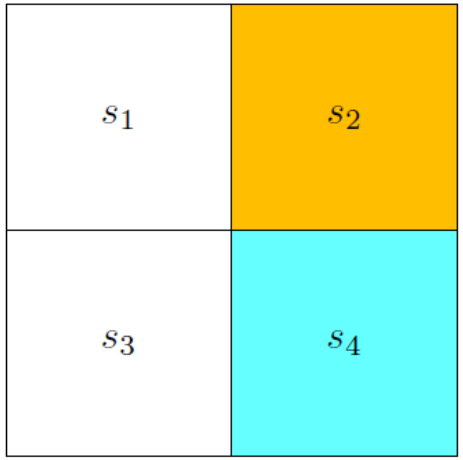


q-table	a_1	a_2	a_3	a_4	a_5
s_1	$-1+\gamma 0$	$-1+\gamma 1$	$0+\gamma 1$	$-1+\gamma 0$	$0+\gamma 0$
s_2	$-1+\gamma 1$	$-1+\gamma 1$	$1+\gamma 1$	$0+\gamma 0$	$-1+\gamma 1$
s_3	$0+\gamma 0$	$1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$
s_4	$-1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$	$1+\gamma 1$

Iteração de valor

Passo 1: Atualização de política

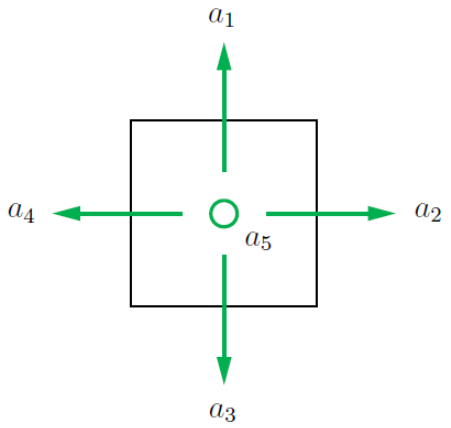
q-table	a_1	a_2	a_3	a_4	a_5
s_1	$-1+\gamma 0$	$-1+\gamma 1$	$0+\gamma 1$	$-1+\gamma 0$	$0+\gamma 0$
s_2	$-1+\gamma 1$	$-1+\gamma 1$	$1+\gamma 1$	$0+\gamma 0$	$-1+\gamma 1$
s_3	$0+\gamma 0$	$1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$
s_4	$-1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$	$1+\gamma 1$



$$\pi_2(a_3|s_1) = 1$$
$$\pi_2(a_3|s_2) = 1$$
$$\pi_2(a_2|s_3) = 1$$
$$\pi_2(a_5|s_4) = 1$$

$$\pi(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases}$$

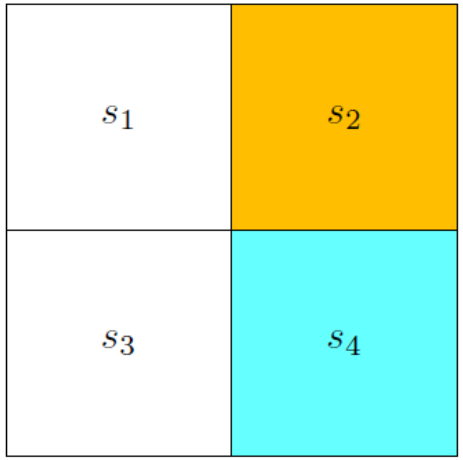
$$onde \quad a_k^*(s) = \underset{a}{\operatorname{argmax}}\, q(s,a)$$



Iteração de valor

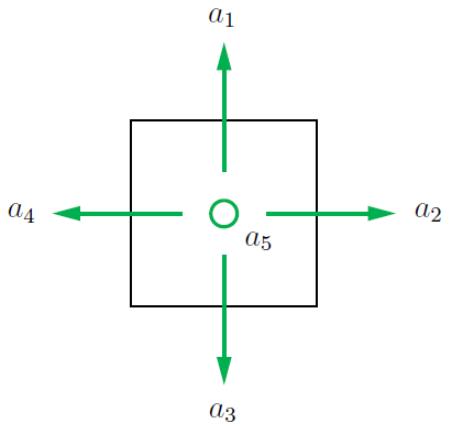
Passo 2: Atualização de valor

q-table	a_1	a_2	a_3	a_4	a_5
s_1	$-1+\gamma 0$	$-1+\gamma 1$	$0+\gamma 1$	$-1+\gamma 0$	$0+\gamma 0$
s_2	$-1+\gamma 1$	$-1+\gamma 1$	$1+\gamma 1$	$0+\gamma 0$	$-1+\gamma 1$
s_3	$0+\gamma 0$	$1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$
s_4	$-1+\gamma 1$	$-1+\gamma 1$	$-1+\gamma 1$	$0+\gamma 1$	$1+\gamma 1$



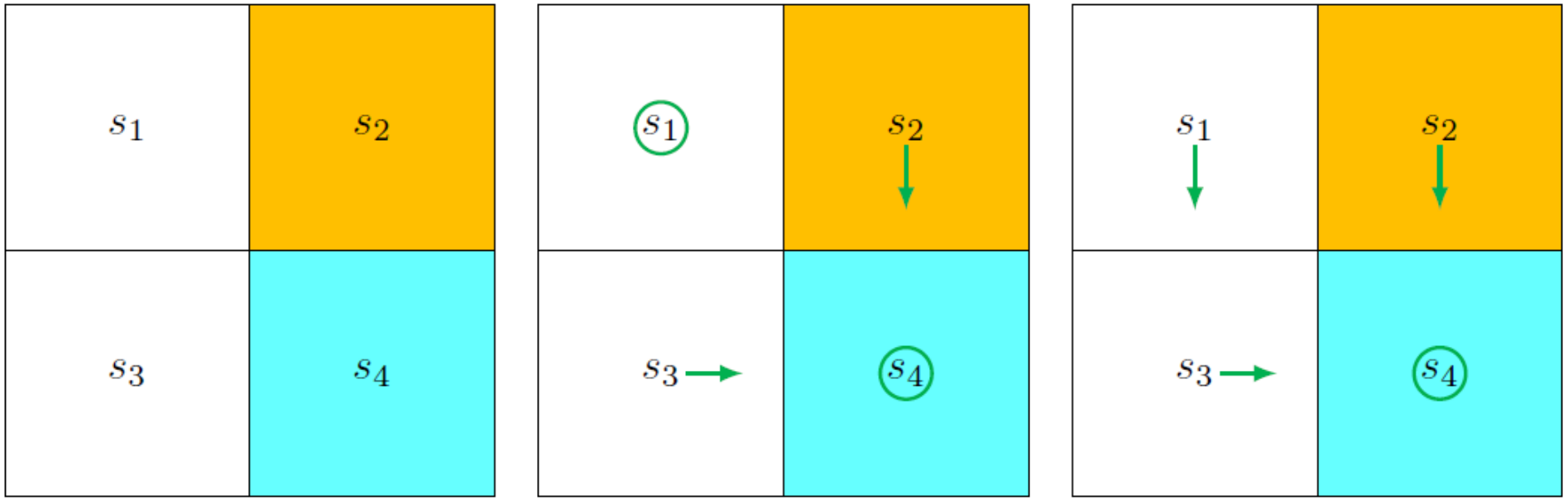
$$v_2(s_1) = \gamma 1$$
$$v_2(s_2) = 1 + \gamma 1$$
$$v_2(s_3) = 1 + \gamma 1$$
$$v_2(s_4) = 1 + \gamma 1$$

$$v_{k+1}(s) = \max_a q_k(s,a)$$



Iteração de valor

- Continua para $k = 2,3,4,5,6 \cdots$
- Pode ser observado que π_2 já é a política ótima.



$$r_{boundary} = r_{forbidden} = -1\,,\, r_{target} = 1,\ \gamma = 0.9$$

Iteração de valor

- Algoritmo

Algorithm 4.1: Value iteration algorithm

Initialization: The probability models $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a) are known. Initial guess v_0.
Goal: Search for the optimal state value and an optimal policy for solving the Bellman optimality equation.

While v_k has not converged in the sense that $\|v_k - v_{k-1}\|$ is greater than a predefined small threshold, for the kth iteration, do

For every state $s \in \mathcal{S}$, do

For every action $a \in \mathcal{A}(s)$, do

q-value: $q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v_k(s')$

Maximum action value: $a_k^*(s) = \arg\max_a q_k(s,a)$

Policy update: $\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise

Value update: $v_{k+1}(s) = \max_a q_k(s,a)$

$$v_k(s) \rightarrow q_k(s) \rightarrow \pi_{k+1}(s)\ gulosa \rightarrow v_{k+1}(s) = \max_a q_k(s,a)$$

Iteração de política

- Iteração de política não resolve diretamente a Equação de otimalidade de Bellman.
- Porém, converge para uma política ótima.
- Iteração de política é um algoritmo iterativo aninhado em outro algoritmo iterativo.
- O algoritmo tem 2 passos
  - Avaliação de política
    - Avalia o valor de estado de uma política dada
  - Melhoria de política
    - Gera uma nova política melhor que a anterior

Iteração de política

- **Passo 1**: Avaliação de política (resolver a equação de Bellman)

$$v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$$

π_k: obtida da iteração anterior

v_{π_k}: valor de estado que queremos calcular

- **Passo 2**: Melhoria de política

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}\left(r_{\pi} + \gamma P_{\pi} v_{\pi_k}\right)$$

π_{k+1}: vai ser utilizado na próxima iteração

Iteração de política

- **Passo 1**: Avaliação de política (resolver a equação de Bellman)

$$v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$$

  - Solução analítica (teórico, não prático)

$$v_{\pi_k} = \left(I - \gamma P_{\pi_k}\right)^{-1} r_{\pi_k}$$

  - Solução iterativa

$$v_{\pi_k}^{(j+1)} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}^{(j)}, \qquad j = 0,1,2, \ldots$$

    - Partindo de um valor inicial $v_{\pi_k}^{(0)}$, é garantido que $v_{\pi_k}^{(j)} \to v_{\pi_k}$ quando $j \to \infty$.

Iteração de política

- **Passo 2**: Melhoria de política
  - Nova Política:

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}\left(r_\pi + \gamma P_\pi v_{\pi_k}\right)$$

  - Lema de melhora da política: (a política resultante é melhor ou igual): se $\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}\left(r_\pi + \gamma P_\pi v_{\pi_k}\right)$, então $v_{\pi_{k+1}}(s) \geq v_{\pi_k}(s)$, para todo $s \in \mathcal{S}$.

Iteração de política

- Sequências geradas:
  1. Políticas: $\{\pi_0, \pi_1, \cdots, \pi_k, \cdots\}$
  2. Valores de estado: $\{v_{\pi_0}, v_{\pi_1}, \cdots, v_{\pi_k}, \cdots\}$
- De acordo com o lema de melhora da política:

$$v_{\pi_0} \leq v_{\pi_1} \leq v_{\pi_2} \leq \cdots \leq v_{\pi_k} \leq \cdots \leq \boxed{v^*}$$

- *Teorema da convergência da iteração de política: A sequência $\{v_{\pi_k}\}_{k=0}^{\infty}$ gerada pelo algoritmo de iteração de política converge para o valor de estado ótimo v^*. Como consequência, a sequência de políticas $\{\pi_k\}_{k=0}^{\infty}$ converge para uma política ótima.*
  - $v_{\pi_k} \rightarrow v^*$ *(valor de estado ótimo)*
  - $\pi_k \rightarrow \pi^*$ *(política ótima)*

Iteração de política

- *Iteração de política vs. Iteração de valor*
  - *Iteração de política requer múltiplas iterações internas (avaliação de política)*
  - *Iteração de política converge em menos iterações externas, porém cada iteração pode ser custosa*
  - *Iteração de valor faz uma atualização mais simples, mas pode precisar de mais iterações para convergir.*

Iteração de política (forma escalar)

- **Passo 1**: Avaliação de política

$$v_{\pi_k}^{(j+1)}(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}^{(j)}(s') \right], \qquad s \in \mathcal{S}, \qquad j = 0,1,2, \ldots$$

- **Passo 2:** Melhoria de política

$$\pi_{k+1}(s) = \operatorname*{argmax}_\pi \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}(s') \right], \qquad s \in \mathcal{S}$$

$$\pi_{k+1}(s) = \operatorname*{argmax}_\pi \sum_a \pi(a|s)\, q_{\pi_k}(s,a), \qquad s \in \mathcal{S}$$

A política ótima (**gulosa**):

$$\boxed{\pi_{k+1}(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases}}, \qquad onde \qquad a_k^*(s) = \operatorname*{argmax}_a q_{\pi_k}(s,a)$$

Iteração de valor (forma escalar)

- **Passo 2**: Atualização de valor

$$v_{k+1}(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right], \qquad s \in \mathcal{S}$$

$$v_{k+1}(s) = \sum_a \pi(a|s) q_k(s,a) , \qquad s \in \mathcal{S}$$

- De acordo com o passo 1:

$$\boxed{v_{k+1}(s) = \max_a q_k(s,a)}, \qquad s \in \mathcal{S}$$

Iteração de política

- Algoritmo

Algorithm 4.2: Policy iteration algorithm

Initialization: The system model, $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a), is known. Initial guess π_0.
Goal: Search for the optimal state value and an optimal policy.

While v_{π_k} has not converged, for the kth iteration, do
Policy evaluation:
Initialization: an arbitrary initial guess $v_{\pi_k}^{(0)}$
While $v_{\pi_k}^{(j)}$ has not converged, for the jth iteration, do
For every state $s \in \mathcal{S}$, do
$$v_{\pi_k}^{(j+1)}(s) = \sum_a \pi_k(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}^{(j)}(s') \right]$$
Policy improvement:
For every state $s \in \mathcal{S}$, do
For every action $a \in \mathcal{A}$, do
$$q_{\pi_k}(s,a) = \sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}(s')$$
$a_k^*(s) = \arg\max_a q_{\pi_k}(s,a)$
$\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise

Referências

- Shiyu Zhao. Mathematical Foundations of Reinforcement Learning. Springer Singapore, 2025. [capítulo 4]
  - 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 4]
  - 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.