Aprendizado por Reforço

Programação dinâmica (parte 2)

Recapitulação das aula passadas...

- Algoritmo:

Iteração de valor

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$

Recapitulação das aula passadas...

- Algoritmo:

Iteração de política

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s)) \left[r + \gamma V(s') \right]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \text{argmax}_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma V(s') \right]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Recapitulação das aula passadas...

- Algoritmo:

Iteração de política

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

Iteração de política (recapitulação da aula passada...)

- **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

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

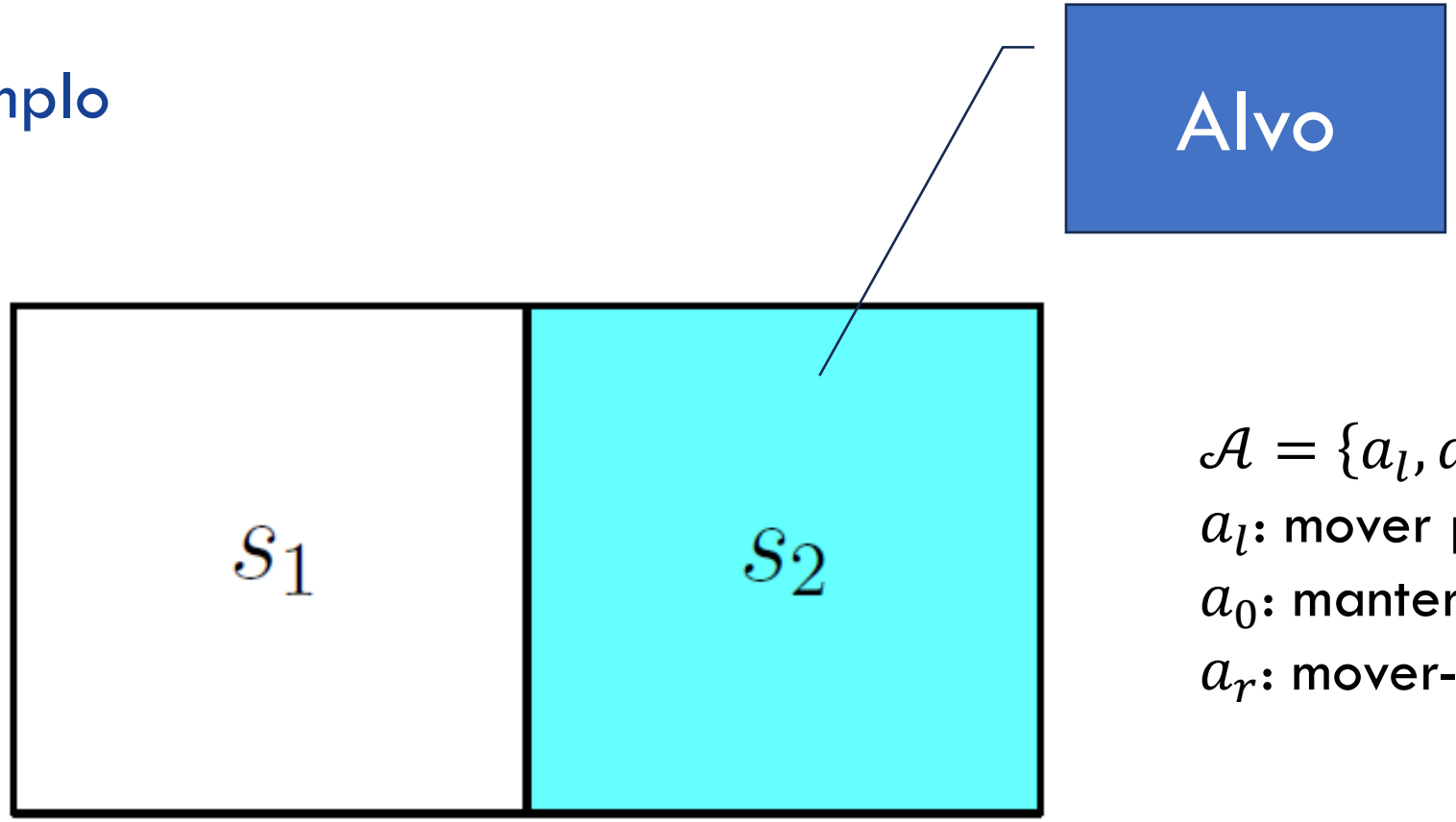
$$\pi_{k+1}(s) = \underset{\pi}{\operatorname{argmax}} \sum_a \pi(a|s)\, q_{\pi_k}(s,a), \qquad s \in \mathcal{S}$$

A política ótima (**gulosa**):

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

Iteração de política

- Exemplo



$\mathcal{A} = \{a_l, a_0, a_r\}$

a_l: mover para esquerda

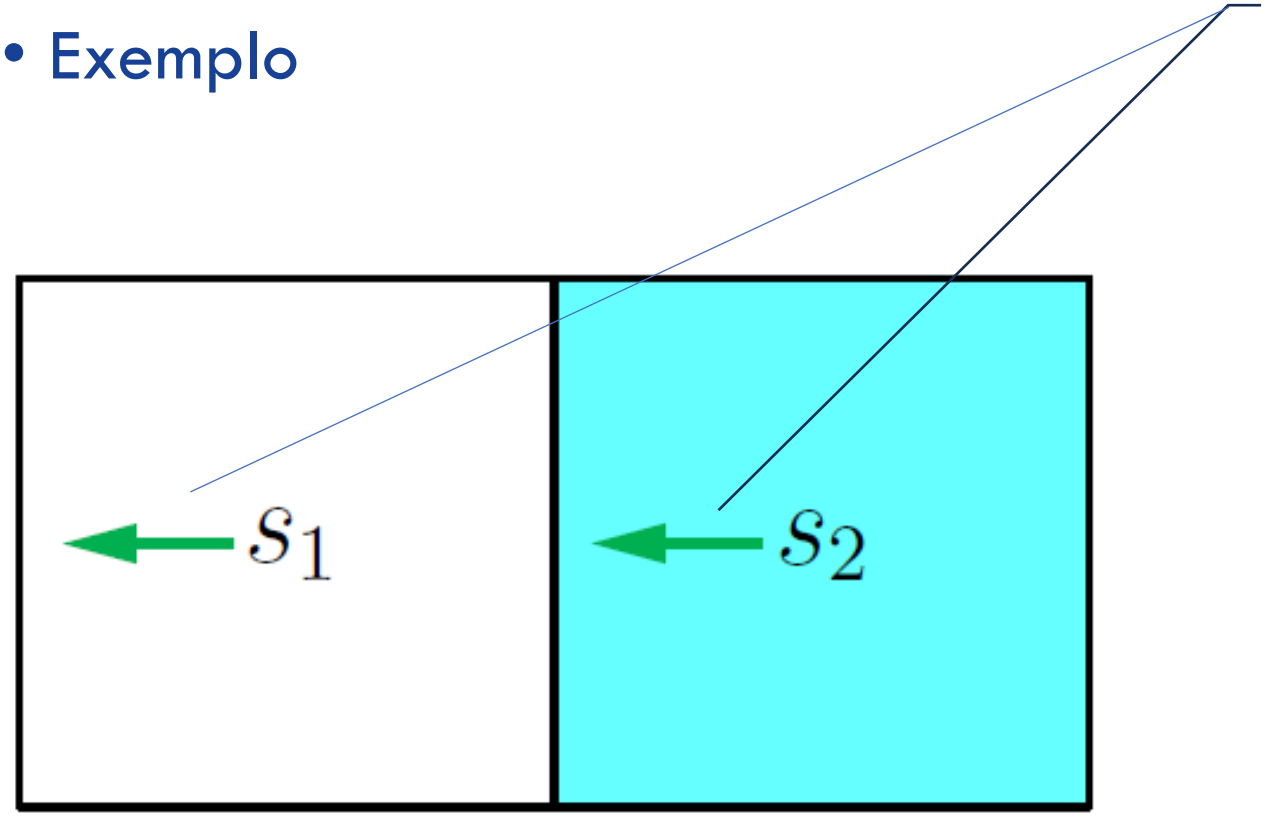
a_0: manter-se no mesmo estado

a_r: mover-se para direita

$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

Política inicial π_0

$k = 0$ (Avaliação de política)

Equação de Bellman

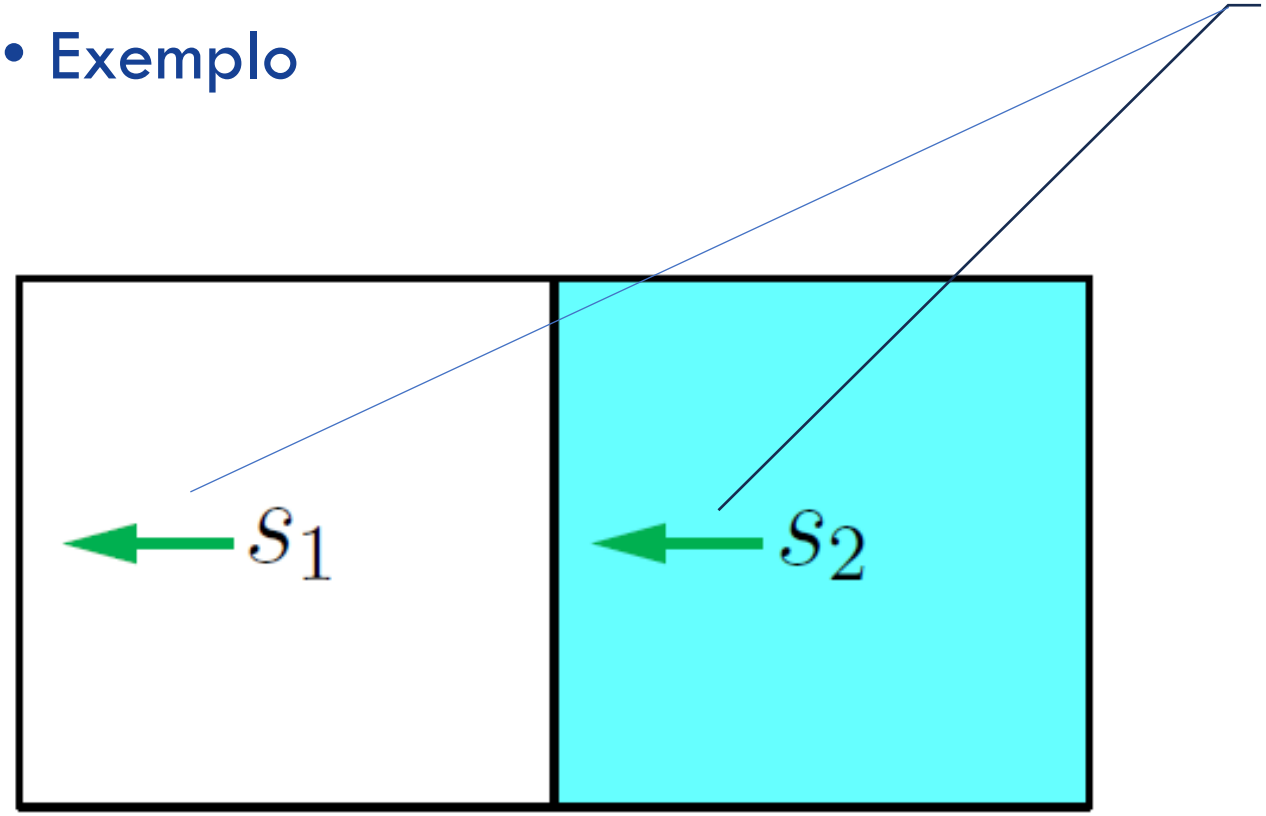
$$v_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

Equação de Bellman para π_0:

$$v_{\pi_0}(s_1) = ?$$
$$v_{\pi_0}(s_2) = ?$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

Política inicial π_0

$k = 0$ (Avaliação de política)

Equação de Bellman

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

Equação de Bellman para π_0:

$$v_{\pi_0}(s_1) = -1 + \gamma v_{\pi_0}(s_1)$$

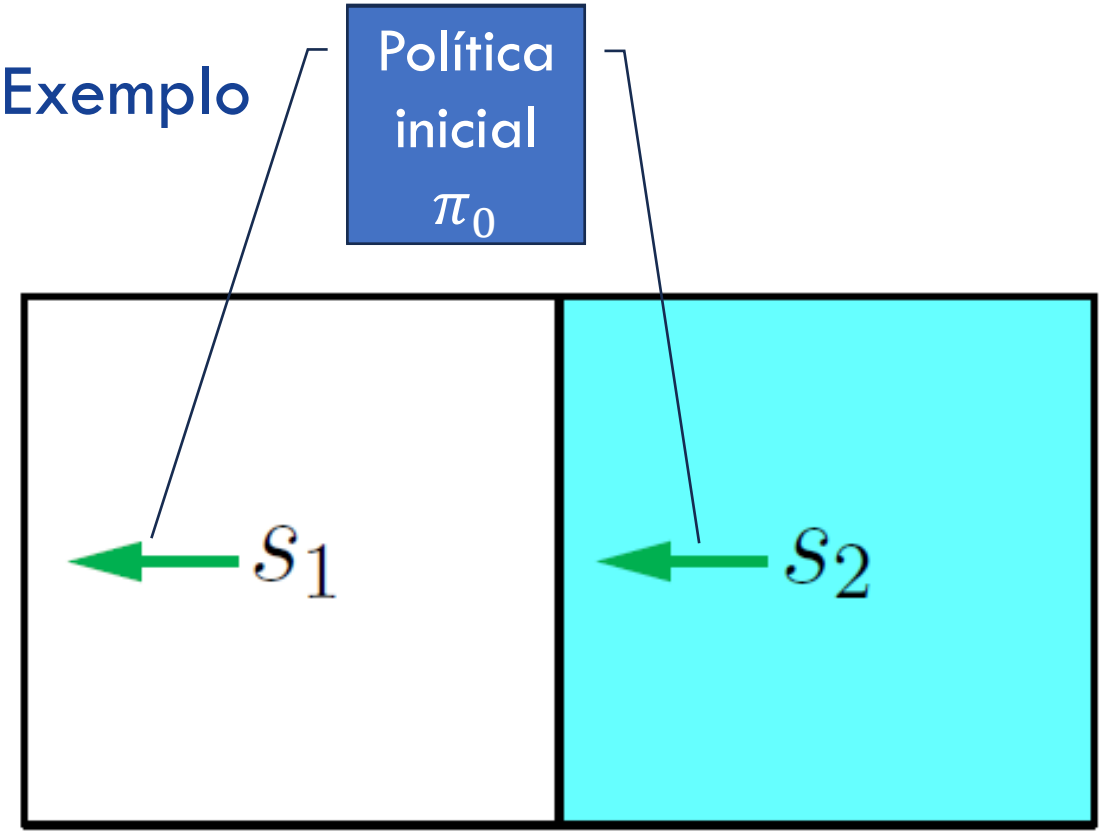
$$v_{\pi_0}(s_2) = 0 + \gamma v_{\pi_0}(s_1)$$

Resolvendo:

$$v_{\pi_0}(s_1) = -10, \qquad v_{\pi_0}(s_2) = -9$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\,, r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Avaliação de política)

Equação de Bellman para π_0:

$$\begin{cases} v_{\pi_0}(s_1) = -1 + \gamma v_{\pi_0}(s_1) \\ v_{\pi_0}(s_2) = 0 + \gamma v_{\pi_0}(s_1) \end{cases}$$

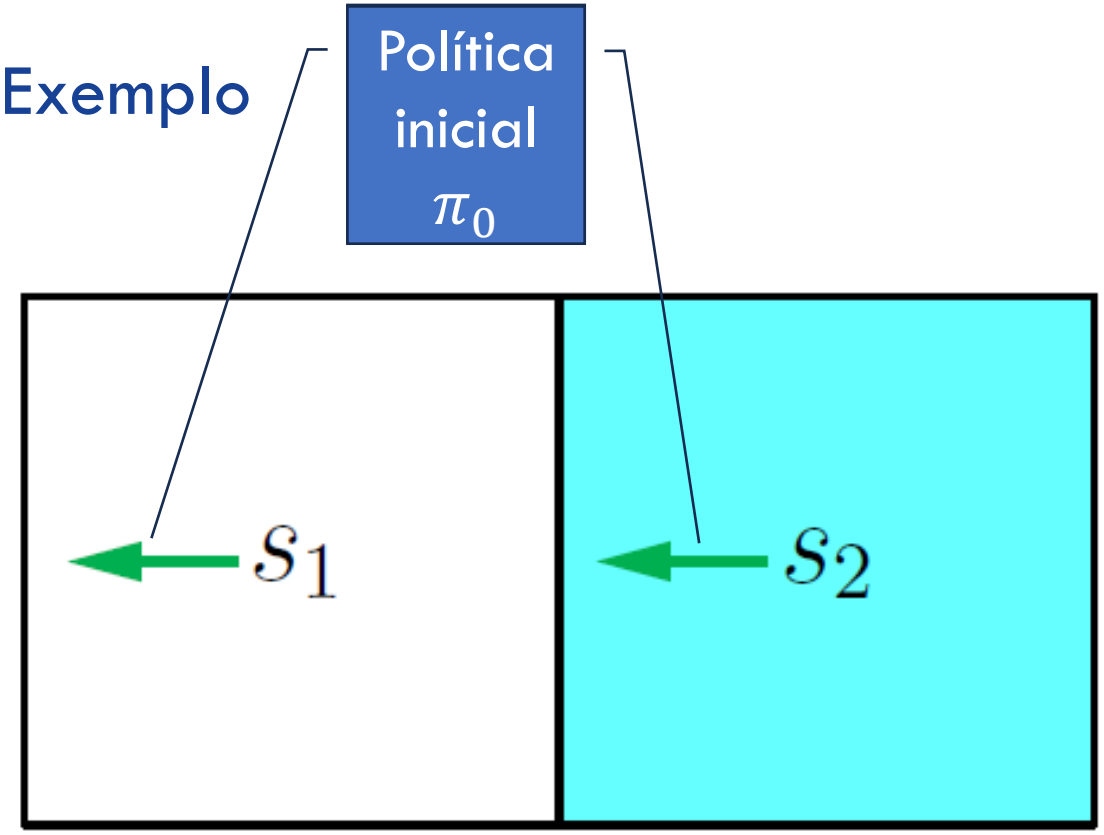
Na prática usamos o método iterativo:

$$v_{\pi_0}^{(0)}(s) \to v_{\pi_0}^{(1)}(s) \to \cdots \to v_{\pi_0}^{(j)}(s) \to \cdots$$

$$\begin{cases} v_{\pi_0}^{(0)}(s_1) = 0 \\ v_{\pi_0}^{(0)}(s_2) = 0 \end{cases}$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\,, r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Avaliação de política)

Equação de Bellman para π_0:

$$\begin{cases} v_{\pi_0}(s_1) = -1 + \gamma v_{\pi_0}(s_1) \\ v_{\pi_0}(s_2) = 0 + \gamma v_{\pi_0}(s_1) \end{cases}$$

Na prática usamos o método iterativo:

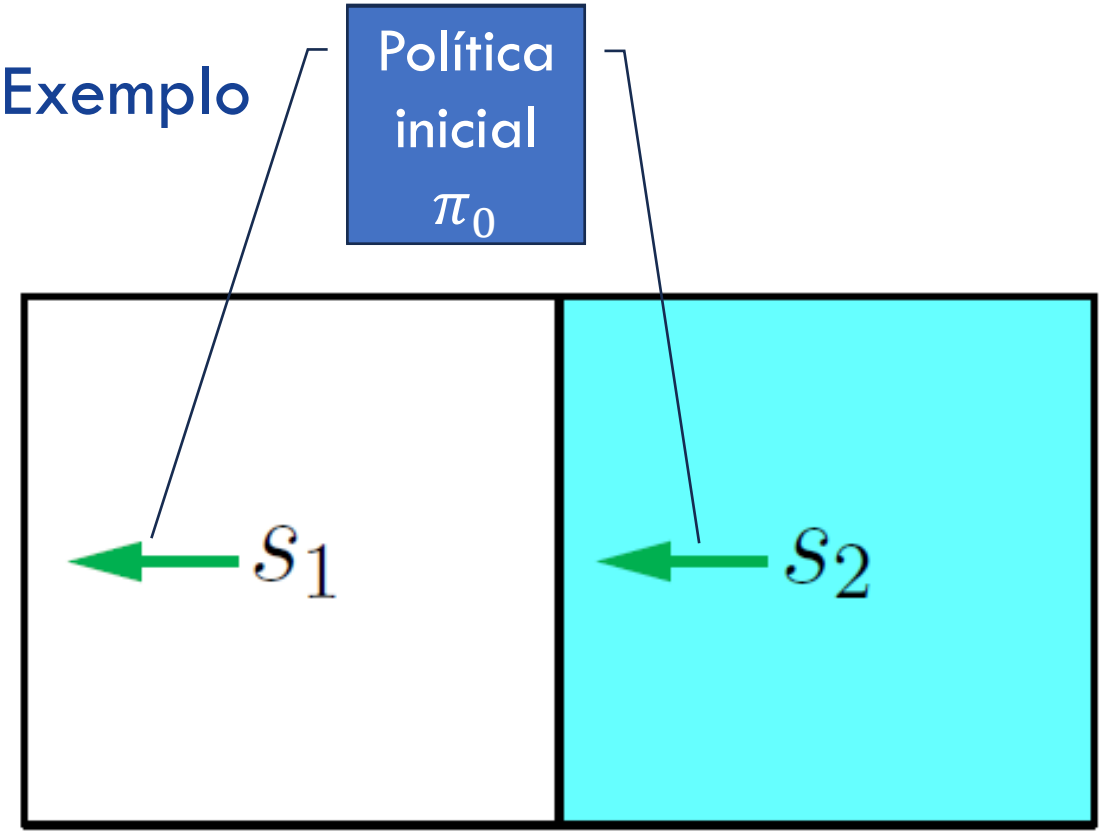
$$v_{\pi_0}^{(0)}(s) \to v_{\pi_0}^{(1)}(s) \to \cdots \to v_{\pi_0}^{(j)}(s) \to \cdots$$

$$\begin{cases} v_{\pi_0}^{(0)}(s_1) = 0 \\ v_{\pi_0}^{(0)}(s_2) = 0 \end{cases}$$

$$\begin{cases} v_{\pi_0}^{(1)}(s_1) = -1 + \gamma v_{\pi_0}^{(0)}(s_1) = -1 \\ v_{\pi_0}^{(1)}(s_2) = 0 + \gamma v_{\pi_0}^{(0)}(s_1) = 0 \end{cases}$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\,,\, r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Avaliação de política)

Equação de Bellman para π_0:

$$\begin{cases} v_{\pi_0}(s_1) = -1 + \gamma v_{\pi_0}(s_1) \\ v_{\pi_0}(s_2) = 0 + \gamma v_{\pi_0}(s_1) \end{cases}$$

Na prática usamos o método iterativo:

$$v_{\pi_0}^{(0)}(s) \to v_{\pi_0}^{(1)}(s) \to \cdots \to v_{\pi_0}^{(j)}(s) \to \cdots$$

$$\begin{cases} v_{\pi_0}^{(0)}(s_1) = 0 \\ v_{\pi_0}^{(0)}(s_2) = 0 \end{cases}$$

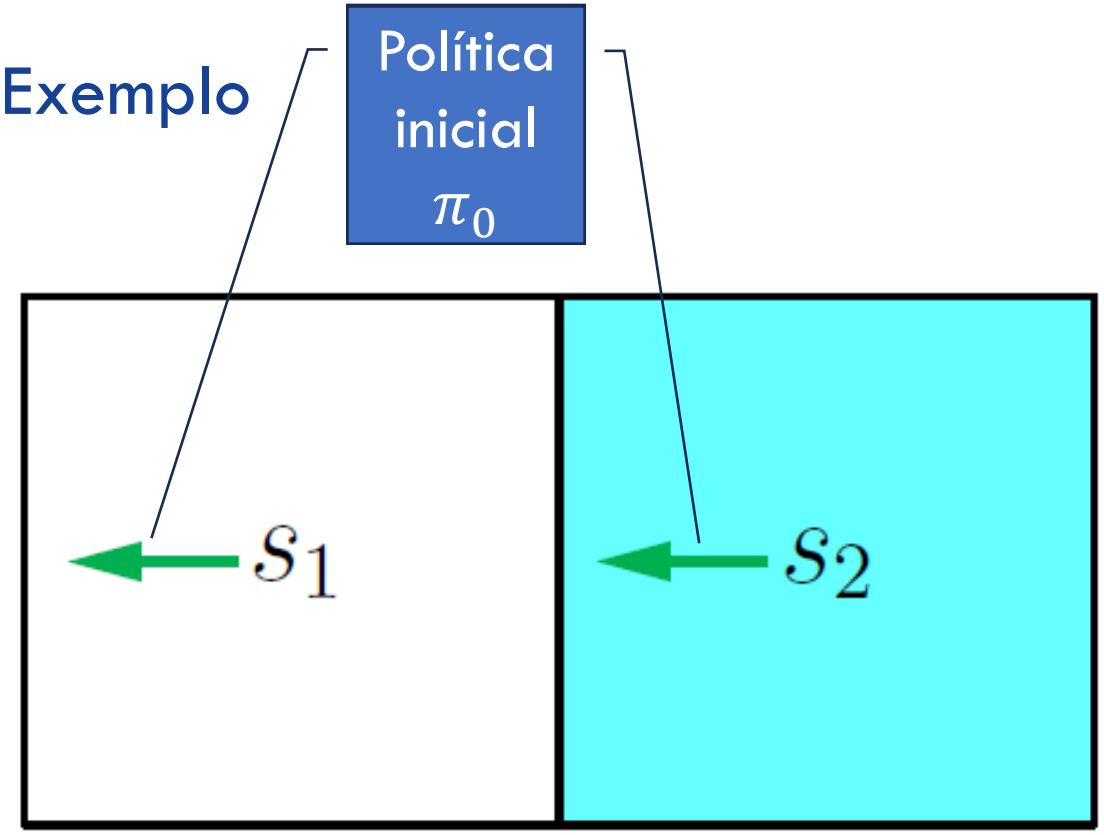
$$\begin{cases} v_{\pi_0}^{(1)}(s_1) = -1 + \gamma v_{\pi_0}^{(0)}(s_1) = -1 \\ v_{\pi_0}^{(1)}(s_2) = 0 + \gamma v_{\pi_0}^{(0)}(s_1) = 0 \end{cases}$$

$$\begin{cases} v_{\pi_0}^{(2)}(s_1) = -1 + \gamma v_{\pi_0}^{(1)}(s_1) = -1.9 \\ v_{\pi_0}^{(2)}(s_2) = 0 + \gamma v_{\pi_0}^{(1)}(s_1) = -0.9 \end{cases}$$

$$\vdots$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Avaliação de política)

Equação de Bellman para π_0:

$$\begin{cases} v_{\pi_0}(s_1) = -1 + \gamma v_{\pi_0}(s_1) \\ v_{\pi_0}(s_2) = 0 + \gamma v_{\pi_0}(s_1) \end{cases}$$

Na prática usamos o método iterativo:

$$v_{\pi_0}^{(0)}(s) \to v_{\pi_0}^{(1)}(s) \to \cdots \to v_{\pi_0}^{(j)}(s) \to \cdots$$

$$\begin{cases} v_{\pi_0}^{(0)}(s_1) = 0 \\ v_{\pi_0}^{(0)}(s_2) = 0 \end{cases}$$

$$\begin{cases} v_{\pi_0}^{(1)}(s_1) = -1 + \gamma v_{\pi_0}^{(0)}(s_1) = -1 \\ v_{\pi_0}^{(1)}(s_2) = 0 + \gamma v_{\pi_0}^{(0)}(s_1) = 0 \end{cases}$$

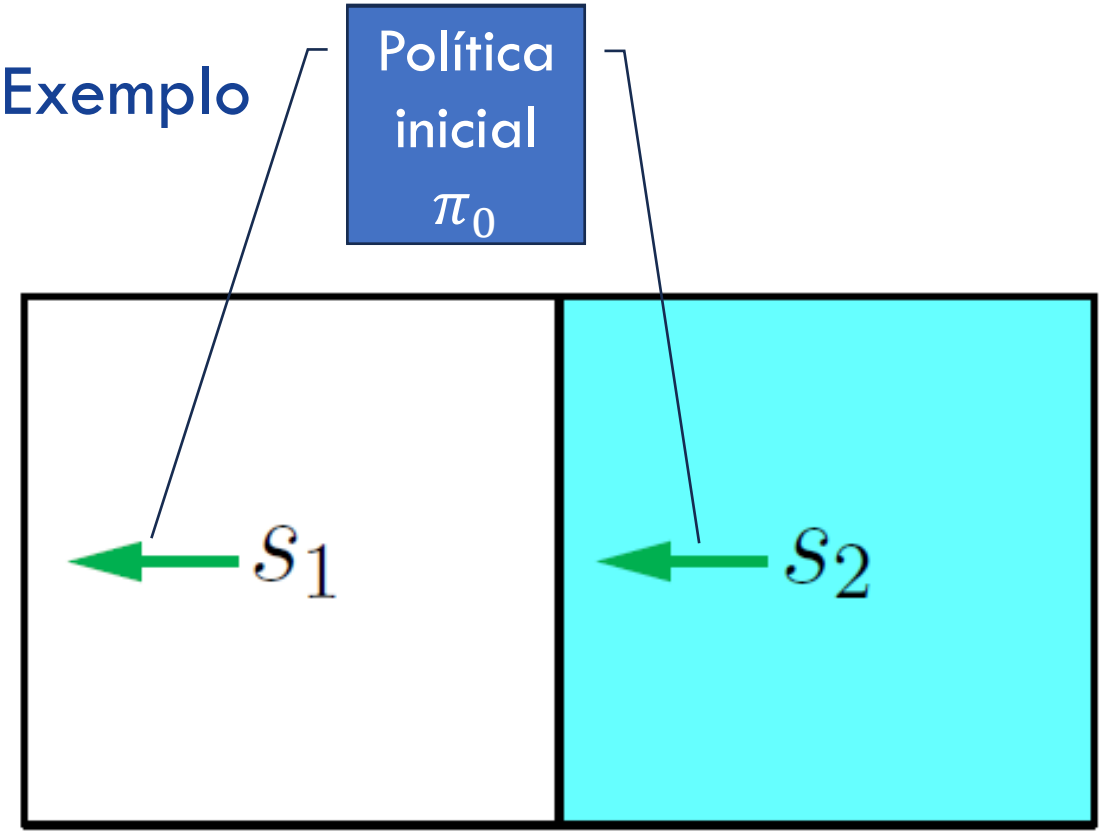
$$\begin{cases} v_{\pi_0}^{(2)}(s_1) = -1 + \gamma v_{\pi_0}^{(1)}(s_1) = -1.9 \\ v_{\pi_0}^{(2)}(s_2) = 0 + \gamma v_{\pi_0}^{(1)}(s_1) = -0.9 \end{cases}$$

$$\vdots$$

$$\begin{cases} v_{\pi_0}^{(j)}(s_1) \to v_{\pi_0}(s_1) = -10 \\ v_{\pi_0}^{(j)}(s_1) \to v_{\pi_0}(s_2) = -9 \end{cases}$$

Iteração de política

- Exemplo



$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Melhoria de política)

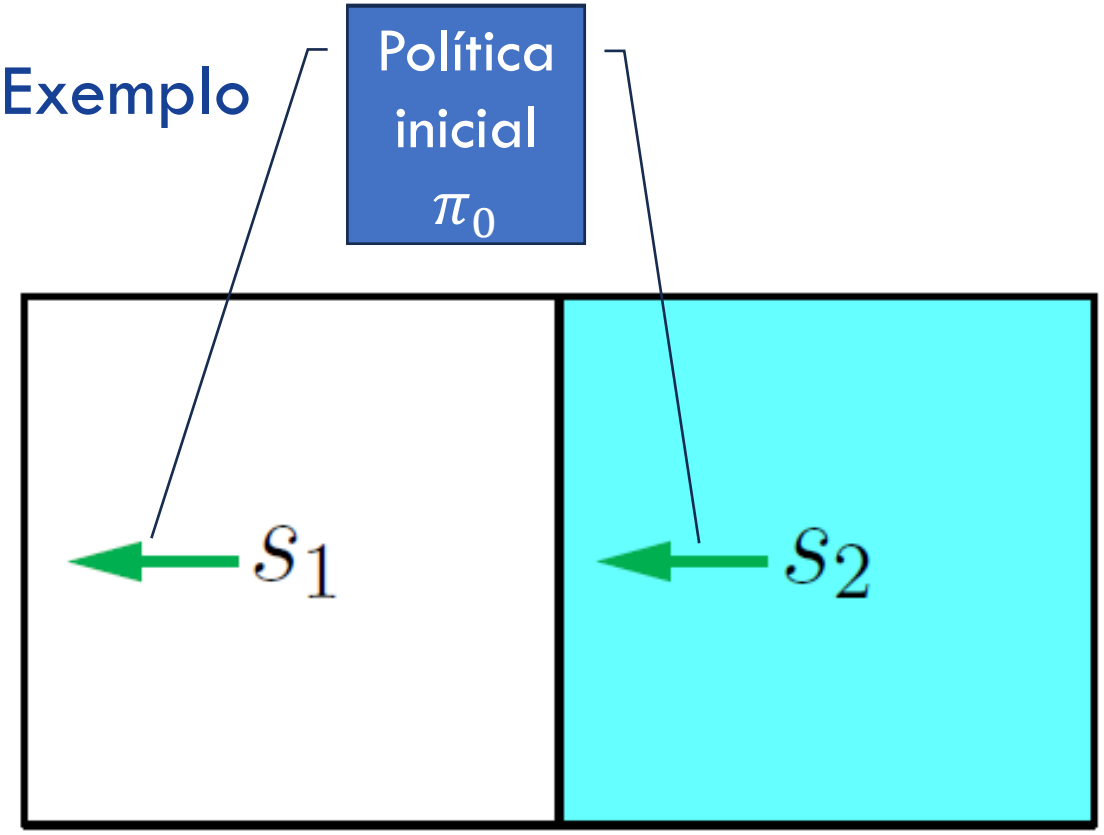
Valor de ação

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

$q_{\pi_k}(s,a)$	a_ℓ	a_0	a_r
s_1			
s_2			

Iteração de política

- Exemplo



$r_{boundary} = -1$, $r_{other} = 0$, $r_{target} = 1$, $\gamma = 0.9$

$k = 0$ (Melhoria de política)

Valor de ação

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

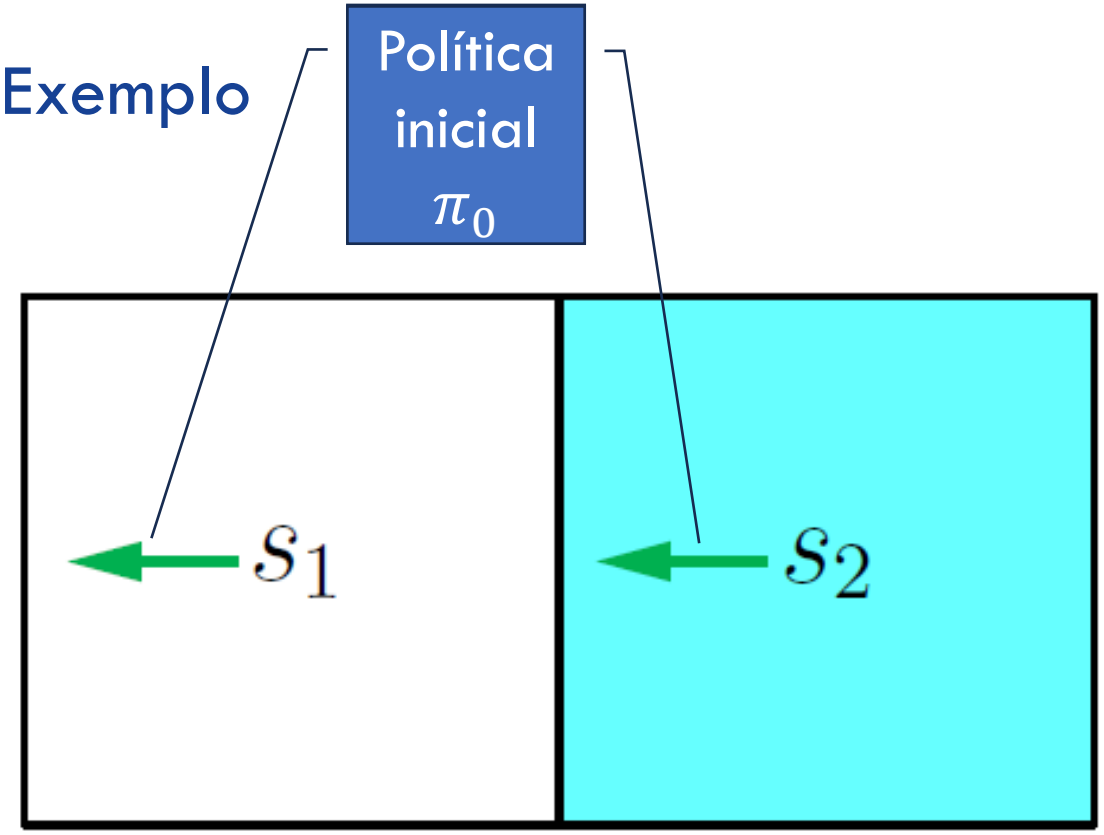
$q_{\pi_k}(s,a)$	a_ℓ	a_0	a_r
s_1	$-1 + \gamma v_{\pi_k}(s_1)$	$0 + \gamma v_{\pi_k}(s_1)$	$1 + \gamma v_{\pi_k}(s_2)$
s_2	$0 + \gamma v_{\pi_k}(s_1)$	$1 + \gamma v_{\pi_k}(s_2)$	$-1 + \gamma v_{\pi_k}(s_2)$

$v_{\pi_0}(s_1) = -10,$
$v_{\pi_0}(s_2) = -9$

$q_{\pi_0}(s,a)$	a_ℓ	a_0	a_r
s_1	-10	-9	-7.1
s_2	-9	-7.1	-9.1

Iteração de política

- Exemplo



$$r_{boundary} = -1\ , r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Melhoria de política)

Valor de ação

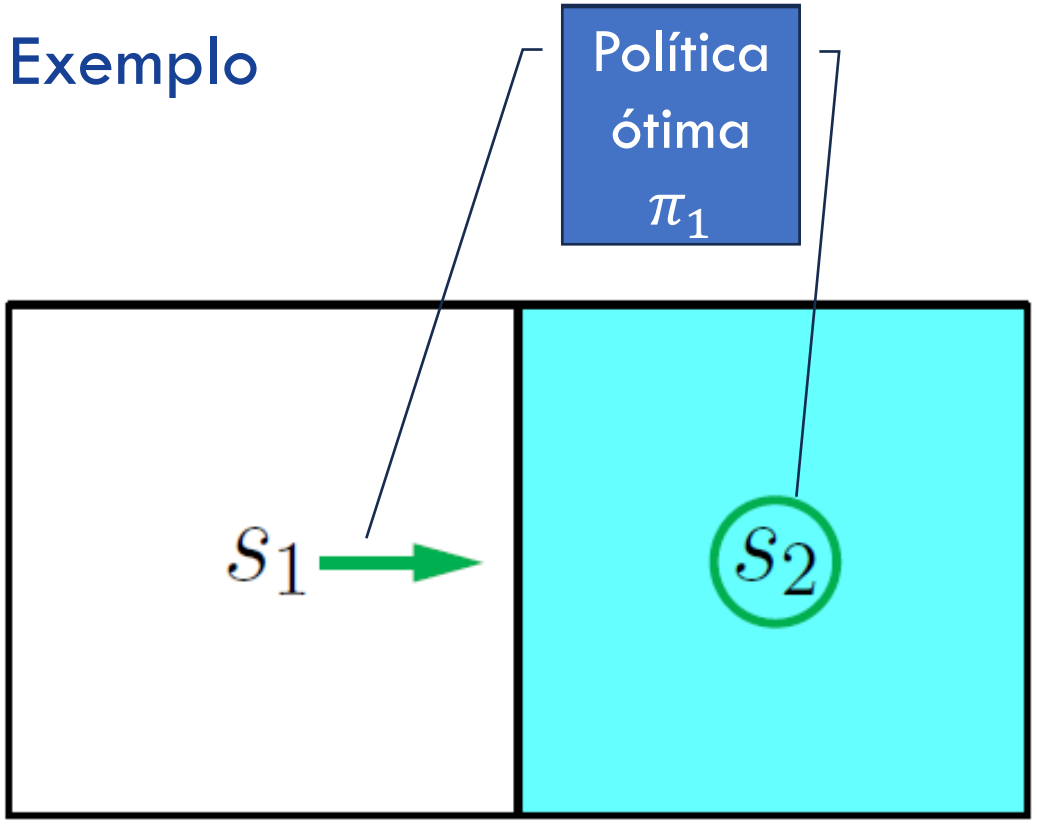
$$q_k(s,a) = \sum_r p(\,r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s')$$

$q_{\pi_k}(s,a)$	a_ℓ	a_0	a_r
s_1	$-1+\gamma v_{\pi_k}(s_1)$	$0+\gamma v_{\pi_k}(s_1)$	$1+\gamma v_{\pi_k}(s_2)$
s_2	$0+\gamma v_{\pi_k}(s_1)$	$1+\gamma v_{\pi_k}(s_2)$	$-1+\gamma v_{\pi_k}(s_2)$

$q_{\pi_0}(s,a)$	a_ℓ	a_0	a_r
s_1	-10	-9	-7.1
s_2	-9	-7.1	-9.1

Iteração de política

- Exemplo



$$r_{boundary} = -1\,, r_{other} = 0, r_{target} = 1,$$
$$\gamma = 0.9$$

$k = 0$ (Melhoria de política)

Valor de ação

$q_{\pi_0}(s,a)$	a_ℓ	a_0	a_r
s_1	-10	-9	-7.1
s_2	-9	-7.1	-9.1

Política ótima (**gulosa**):

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

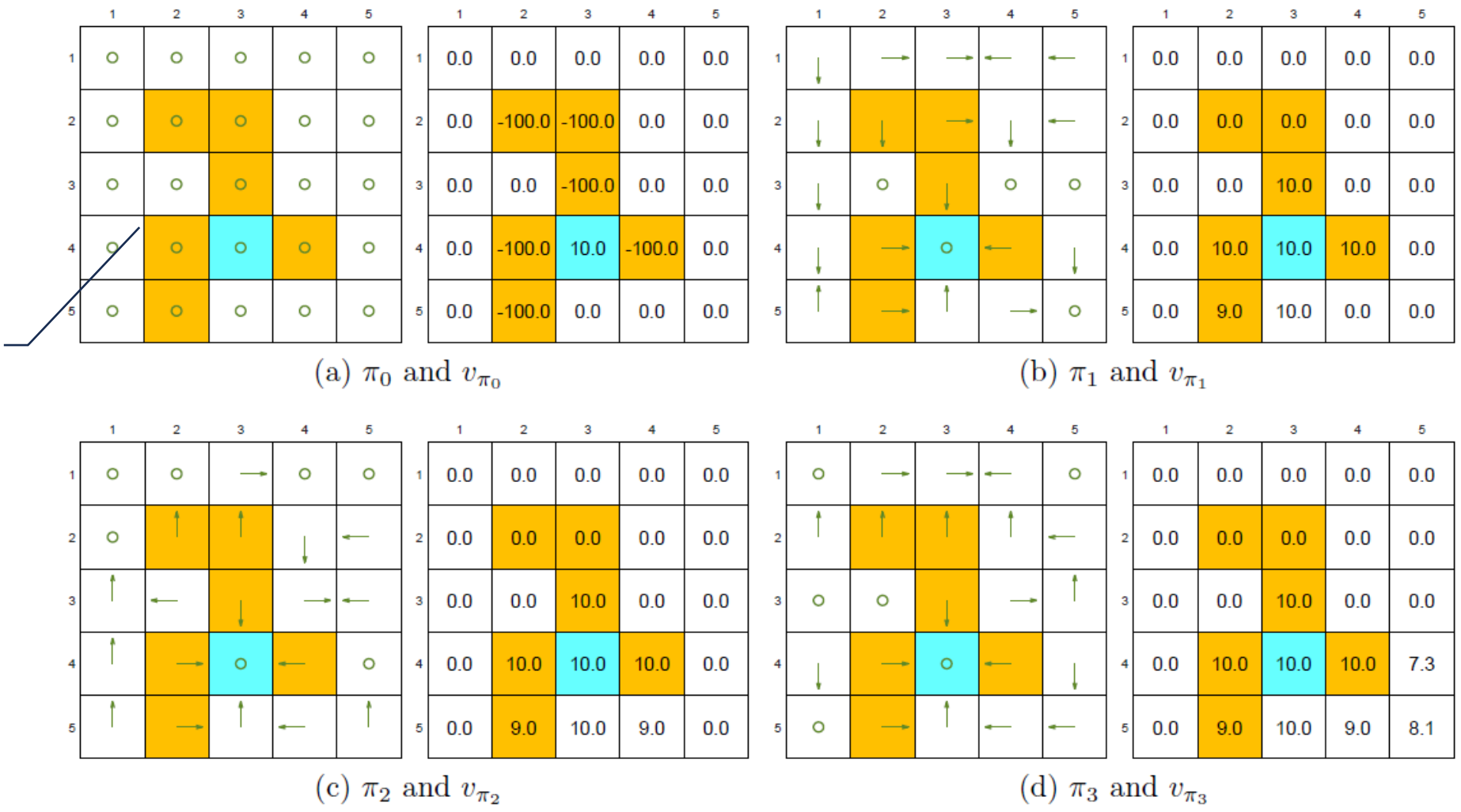
$$\pi_1(a_r|s_1) = 1, \qquad \pi_1(a_0|s_2) = 1$$

Iteração de política

- Exemplo

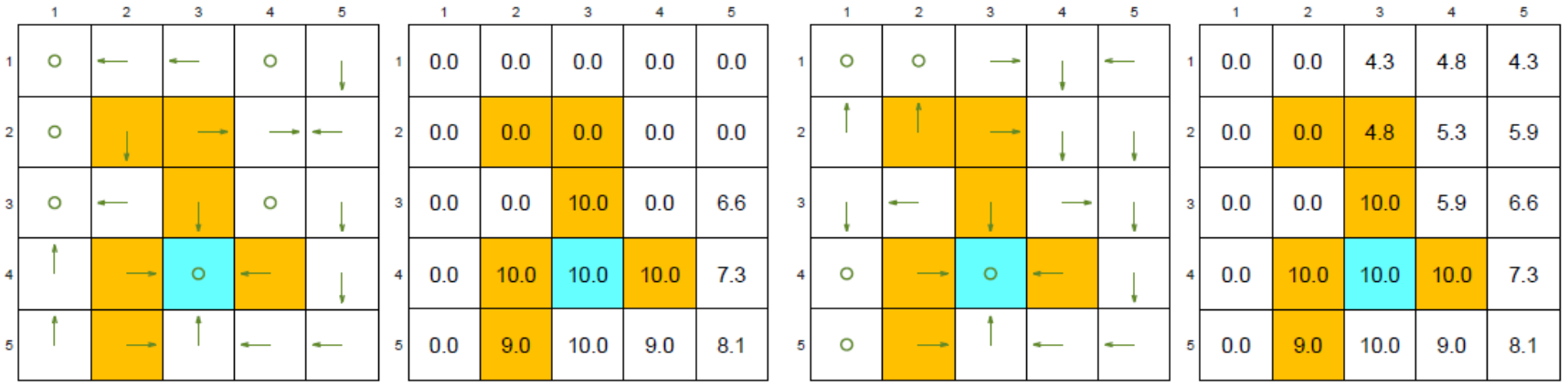
$$r_{boundary} = -1\ ,\ r_{forbidden} = -1\ r_{other} = 0, r_{target} = 1,\ \gamma = 0.9$$

Política inicial π_0 aleatória

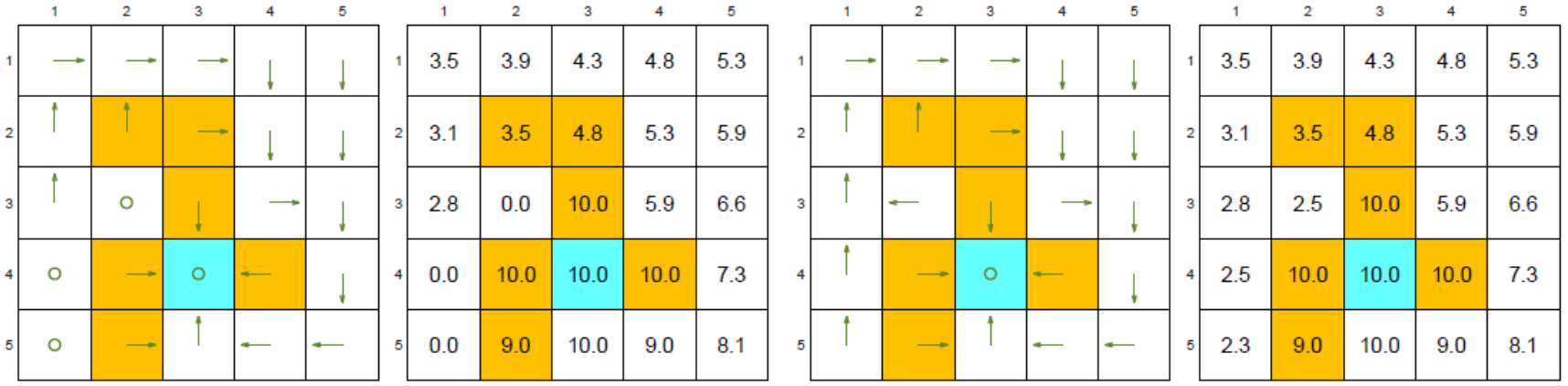


Iteração de política

- Exemplo



(e) π_4 and v_{π_4} (f) π_5 and v_{π_5}

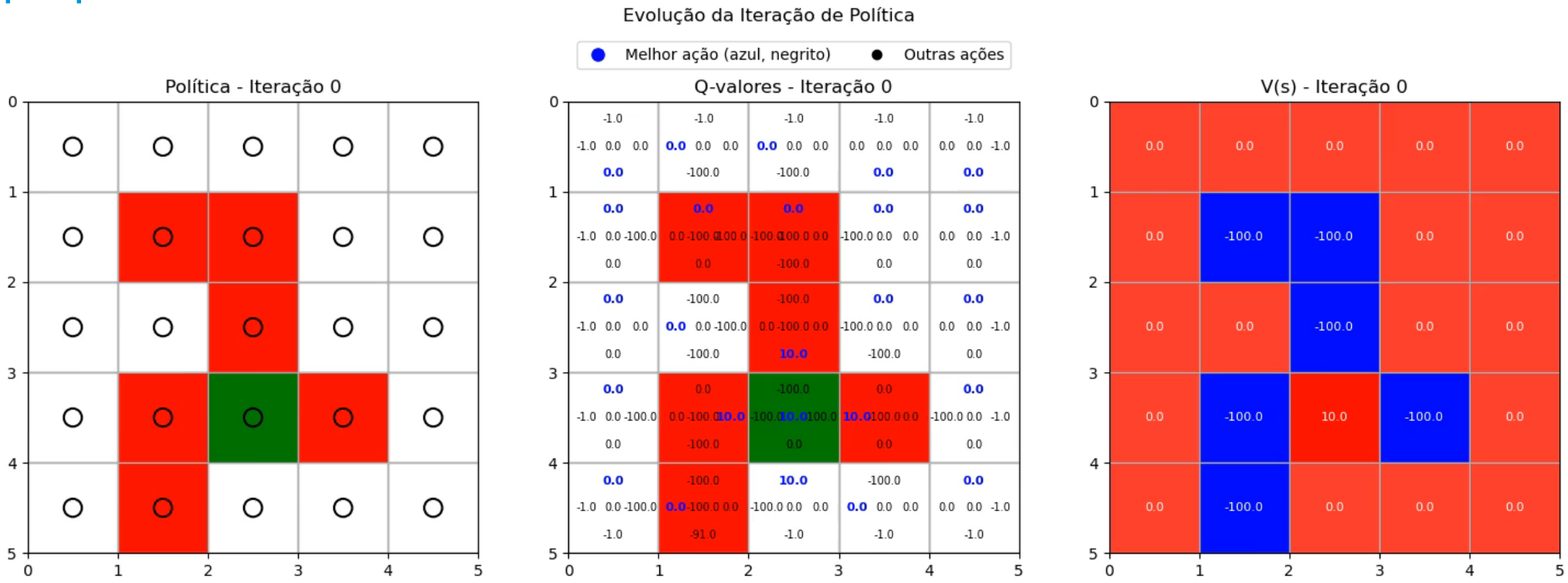


(g) π_9 and v_{π_9} (h) π_{10} and $v_{\pi_{10}}$

$r_{boundary} = -1$
$r_{forbidden} = -1$
$r_{other} = 0$
$r_{target} = 1$
$\gamma = 0.9$

Iteração de política

- Exemplo
  - O que podemos observar?



Iteração de política

- O que podemos observar?
  - O algoritmo de iteração de política consegue convergir para a política ótima mesmo a partir de uma política inicial aleatória.
  - Fenômenos
  1. Estados próximos ao alvo descobrem as trajetórias ótimas mais cedo.
     - Apenas após esses estados próximos encontrarem o caminho ao alvo, os estados mais distantes conseguem planejar caminhos passando por eles.
  2. Estados mais próximos do alvo apresentam valores maiores.
     - Estados distantes exigem mais passos para alcançar recompensa. Devido ao fator de desconto, essas recompensas são reduzidas progressivamente.

Iteração de Política vs. Iteração de Valor

- Iteração de Política
  - Seleciona uma política arbitrária inicial π_0
  - Para cada iteração k:
    - Avaliação de Política (PE)

    Dado π_k , encontrar v_{π_k}

    $v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$
    - Melhoria de Política (PI)

    Dado v_{π_k}, encontrar π_{k+1}

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

- Iteração de Valor:
  - Seleciona um valor de estado arbitrário inicial v_0
  - Para cada iteração k:
    - Atualização de Política (PU)

    Dado v_k, encontrar π_{k+1}

    $\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}(r_{\pi} + \gamma P_{\pi} v_k)$
    - Atualização de Valor (VU)

    Dado π_{k+1}, encontrar v_{k+1}

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

Iteração de Política vs. Iteração de Valor

- Iteração de Política
  - Seleciona uma política arbitrária inicial π_0
  - Para cada iteração k:
    - Avaliação de Política (PE)

    Dado π_k , encontrar v_{π_k}

    $v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$
    - Melhoria de Política (PI)

    Dado v_{π_k}, encontrar π_{k+1}

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

- Iteração de Valor:
  - Seleciona um valor de estado arbitrário inicial v_0
  - Para cada iteração k:
    - Atualização de Política (PU)

    Dado v_k, encontrar π_{k+1}

    $\pi_{k+1} = \underset{\pi}{\operatorname{argmax}}(r_\pi + \gamma P_\pi v_k)$
    - Atualização de Valor (VU)

    Dado π_{k+1}, encontrar v_{k+1}

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

Iteração de política: $\pi_0 \xrightarrow{PE} v_{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} v_{\pi_1} \xrightarrow{PI} \pi_2 \xrightarrow{PE} v_{\pi_2} \xrightarrow{PE} \cdots$

Iteração de valor: $v_0 \xrightarrow{PU} \pi'_1 \xrightarrow{VU} v_1 \xrightarrow{PU} \pi'_2 \xrightarrow{VU} v_2 \xrightarrow{PU} \cdots$

Iteração de Política vs. Iteração de Valor

- Considerando a mesma condição inicial $v_0 = v_{\pi_0}$

	Policy iteration algorithm	Value iteration algorithm	Comments
1) Policy:	π_0	N/A	?
2) Value:	$v_{\pi_0} = r_{\pi_0} + \gamma P_{\pi_0} v_{\pi_0}$	$v_0 \doteq v_{\pi_0}$	?
3) Policy:	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_0})$	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_0)$	?
4) Value:	$v_{\pi_1} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}$	$v_1 = r_{\pi_1} + \gamma P_{\pi_1} v_0$	?
5) Policy:	$\pi_2 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_1})$	$\pi_2' = \arg\max_\pi (r_\pi + \gamma P_\pi v_1)$	?
⋮	⋮	⋮	⋮

Iteração de Política vs. Iteração de Valor

- Considerando a mesma condição inicial $v_0 = v_{\pi_0}$

Resultados idênticos para os 2 algoritmos

Resultados diferentes para os 2 algoritmos

	Policy iteration algorithm	Value iteration algorithm	Comments
1) Policy:	π_0	N/A	
2) Value:	$v_{\pi_0} = r_{\pi_0} + \gamma P_{\pi_0} v_{\pi_0}$	$v_0 \doteq v_{\pi_0}$	
3) Policy:	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_0})$	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_0)$	The two policies are the same
4) Value:	$v_{\pi_1} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}$	$v_1 = r_{\pi_1} + \gamma P_{\pi_1} v_0$	$v_{\pi_1} \geq v_1$ since $v_{\pi_1} \geq v_{\pi_0}$
5) Policy:	$\pi_2 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_1})$	$\pi_2' = \arg\max_\pi (r_\pi + \gamma P_\pi v_1)$	
⋮	⋮	⋮	⋮

Iteração de Política vs. Iteração de Valor

- Considerando a mesma condição inicial $v_0 = v_{\pi_0} = v_{\pi_1}^{(0)}$

	Policy iteration algorithm	Value iteration algorithm
1) Policy:	π_0	N/A
2) Value:	$v_{\pi_0} = r_{\pi_0} + \gamma P_{\pi_0} v_{\pi_0}$	$v_0 \doteq v_{\pi_0}$
3) Policy:	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_0})$	$\pi_1 = \arg\max_\pi (r_\pi + \gamma P_\pi v_0)$
4) Value:	$v_{\pi_1} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}$	$v_1 = r_{\pi_1} + \gamma P_{\pi_1} v_0$
5) Policy:	$\pi_2 = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_1})$	$\pi_2' = \arg\max_\pi (r_\pi + \gamma P_\pi v_1)$
⋮	⋮	⋮

$$
\begin{aligned}
& & v_{\pi_1}^{(0)} &= v_0 \\
\text{value iteration} \leftarrow v_1 \longleftarrow & & v_{\pi_1}^{(1)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(0)} \\
& & v_{\pi_1}^{(2)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(1)} \\
& & &\vdots \\
& & v_{\pi_1}^{(j)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(j-1)} \\
& & &\vdots \\
\text{policy iteration} \leftarrow v_{\pi_1} \longleftarrow & & v_{\pi_1}^{(\infty)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(\infty)}
\end{aligned}
$$

Iteração de Política Truncada

- Uma generalização dos algoritmos de:
  - Iteração de valor
  - Iteração de política
- Ambos podem ser vistos como casos extremos da iteração de política truncada

$$
\begin{aligned}
& & v_{\pi_1}^{(0)} &= v_0 \\
\text{value iteration} \leftarrow v_1 \longleftarrow & & v_{\pi_1}^{(1)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(0)} \\
& & v_{\pi_1}^{(2)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(1)} \\
& & &\vdots \\
\text{truncated policy iteration} \leftarrow \bar{v}_1 \longleftarrow & & v_{\pi_1}^{(j)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(j-1)} \\
& & &\vdots \\
\text{policy iteration} \leftarrow v_{\pi_1} \longleftarrow & & v_{\pi_1}^{(\infty)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(\infty)}
\end{aligned}
$$

Iteração de Política Truncada

- Executa a avaliação de política com número finito de iterações.
  - $j = 1$:
    Iteração de Valor
  - $j \to \infty$:
    Iteração de Política
  - $1 < j = j_{truncado} < \infty$:
    Iteração de política truncada

$$
\begin{aligned}
& & v_{\pi_1}^{(0)} &= v_0 \\
\text{value iteration} \leftarrow v_1 \longleftarrow & & v_{\pi_1}^{(1)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(0)} \\
& & v_{\pi_1}^{(2)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(1)} \\
& & &\vdots \\
\text{truncated policy iteration} \leftarrow \bar{v}_1 \longleftarrow & & v_{\pi_1}^{(j)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(j-1)} \\
& & &\vdots \\
\text{policy iteration} \leftarrow v_{\pi_1} \longleftarrow & & v_{\pi_1}^{(\infty)} &= r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(\infty)}
\end{aligned}
$$

Iteração de política truncada

- Algoritmo

Algorithm 4.3: Truncated policy iteration algorithm

Initialization: The probability models $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a) are known. Initial guess π_0.

Goal: Search for the optimal state value and an optimal policy.

While v_k has not converged, for the kth iteration, do

Policy evaluation:

Initialization: select the initial guess as $v_k^{(0)} = v_{k-1}$. The maximum number of iterations is set as j_{truncate}.

While $j < j_{\text{truncate}}$, do

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

$$v_k^{(j+1)}(s) = \sum_a \pi_k(a|s) \left[\sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_k^{(j)}(s') \right]$$

Set $v_k = v_k^{(j_{\text{truncate}})}$

Policy improvement:

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

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

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

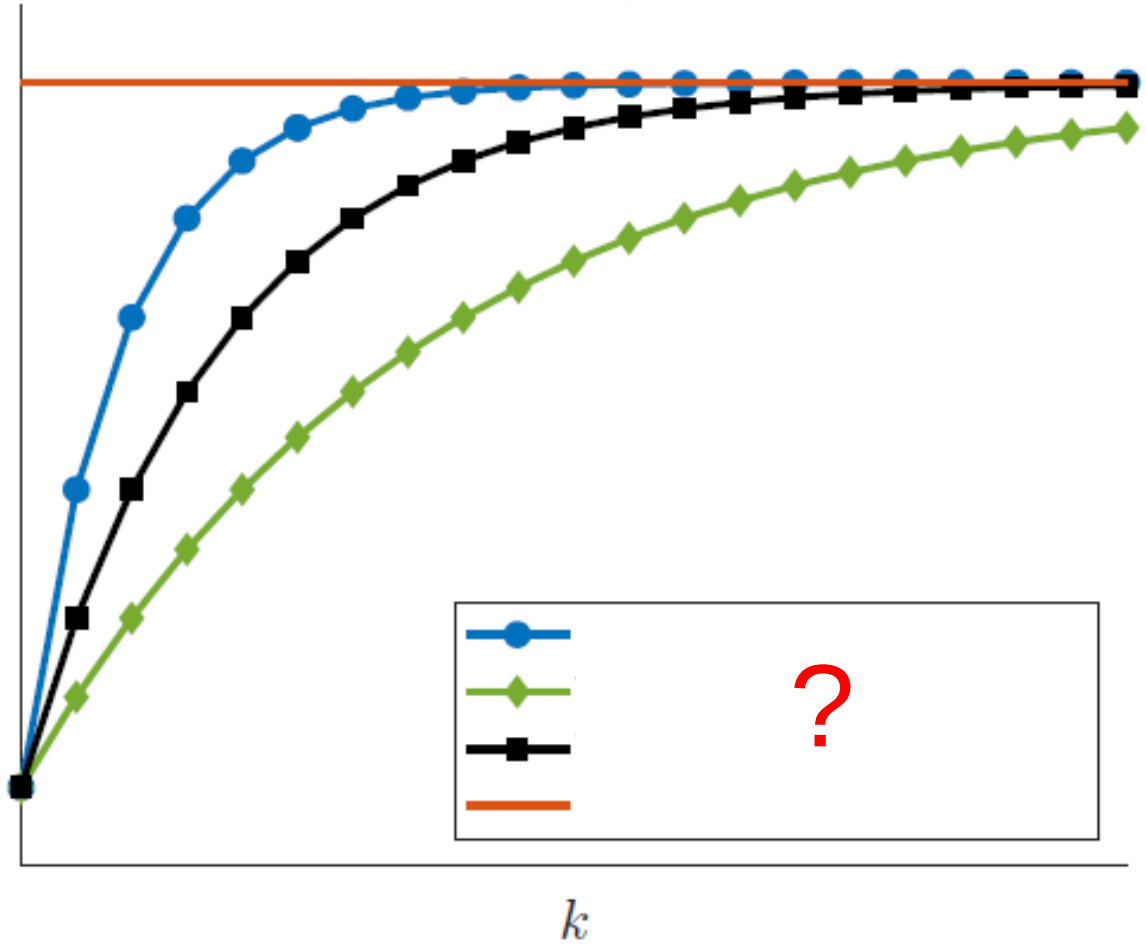
$$a_k^*(s) = \arg\max_a q_k(s,a)$$

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

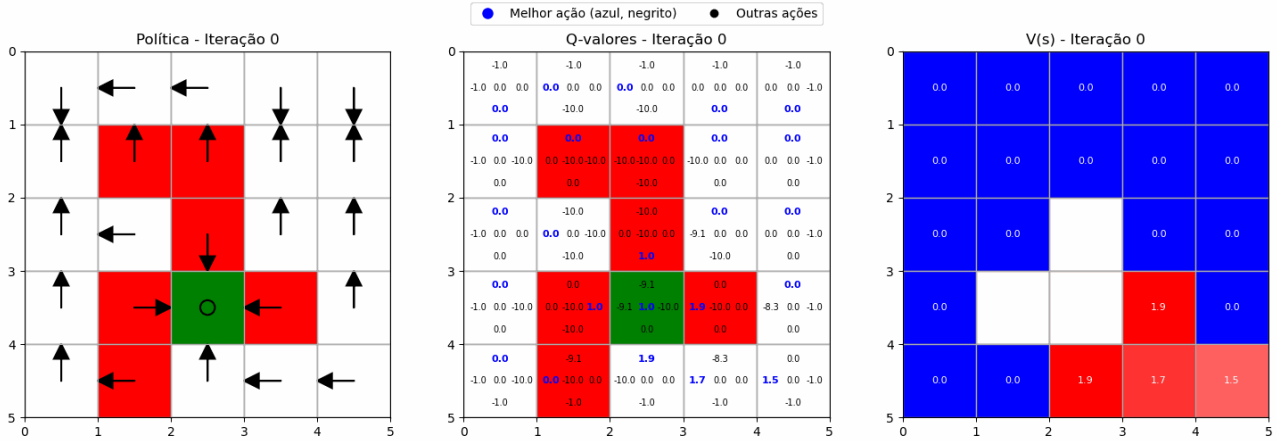
Observação: $v_k^{(j)}$ não são valores de estado, mas sim aproximações.

Iteração de política truncada

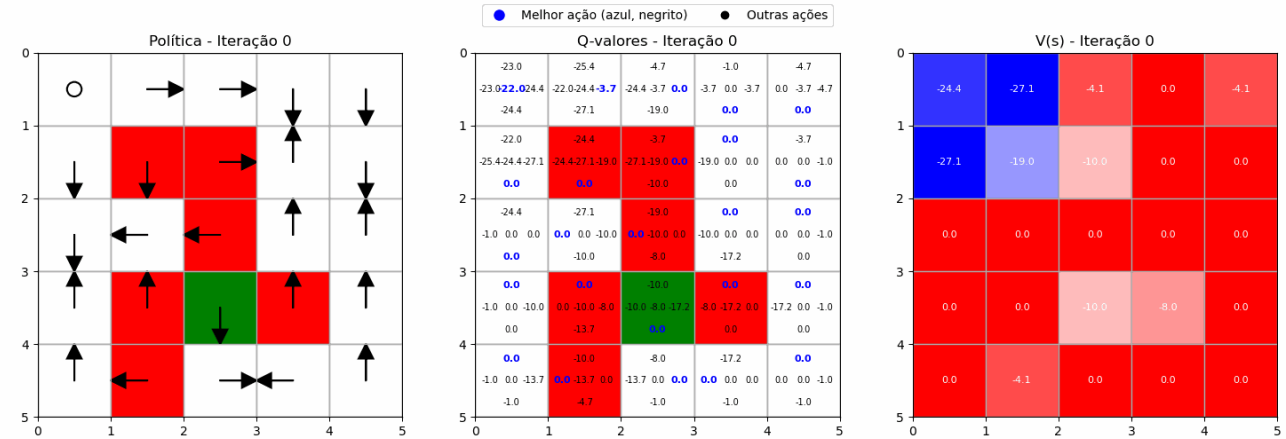
- Convergência



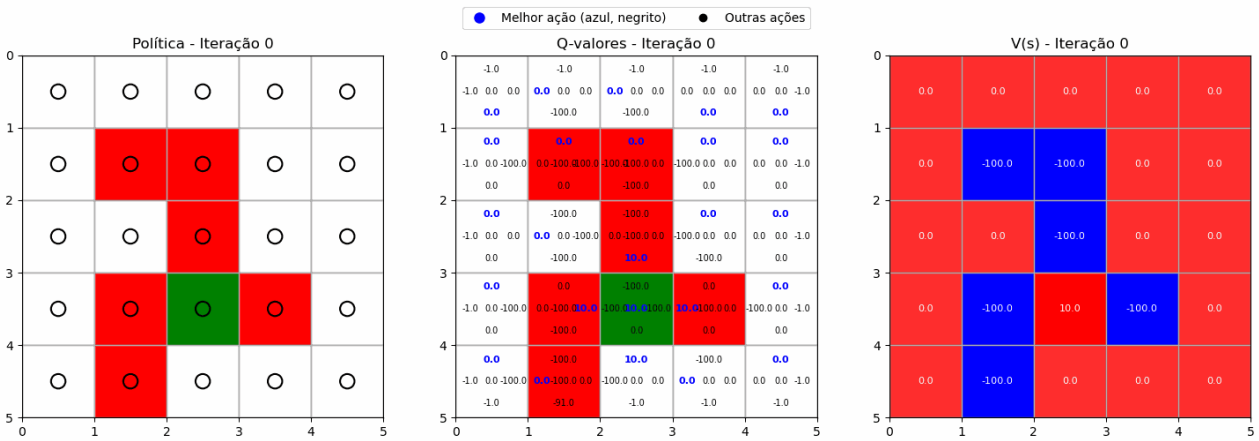
Evolução da Iteração de Valor



Evolução da Política Truncada

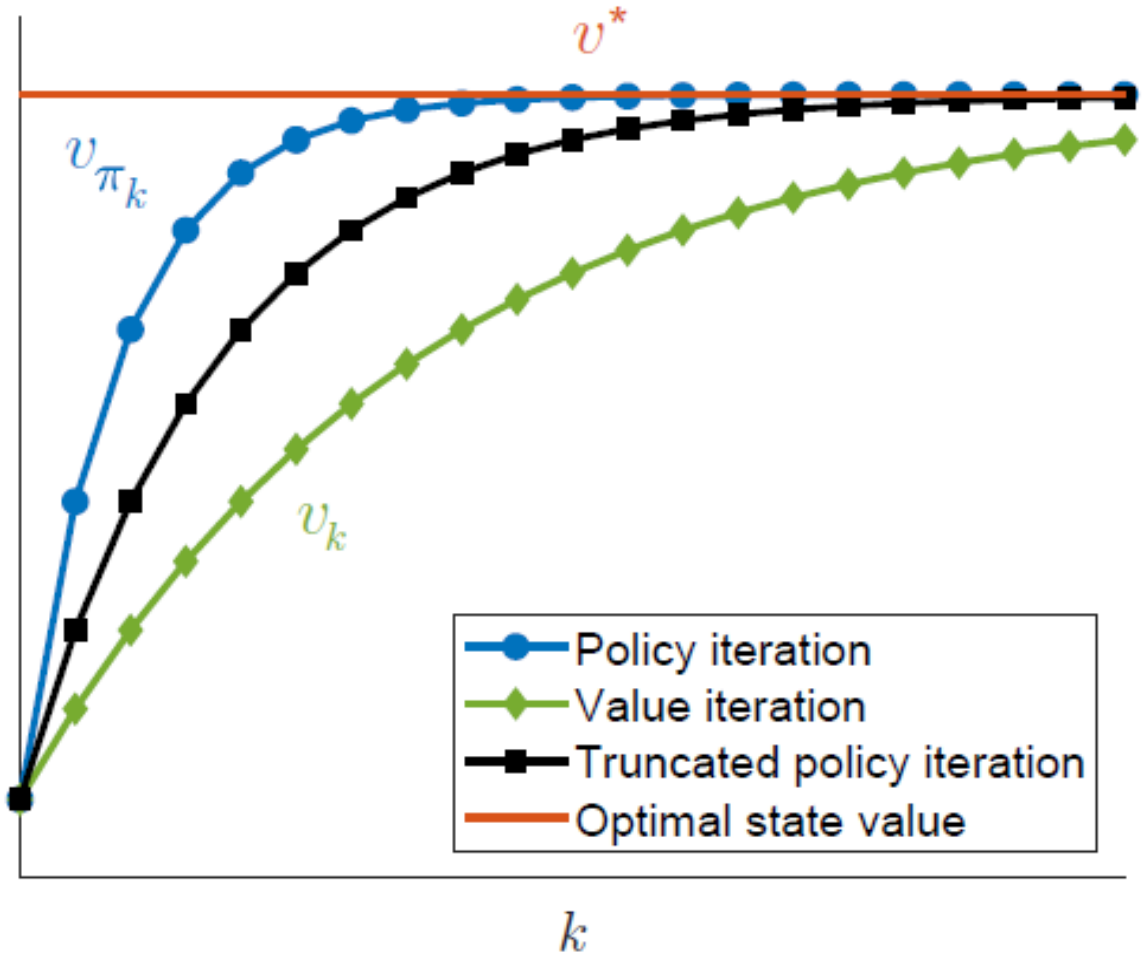


Evolução da Iteração de Política



Iteração de política truncada

- Convergência
  - Mais rápido que a iteração de valor (calcula mais de um iteração durante a avaliação de política)
  - Mais devagar que a iteração de política (calcula apenas um número finito de iterações)



Iteração de política truncada

- Vantagens
  - Mais eficiente computacionalmente que a Iteração de Política (usa menos iterações na avaliação de política)
  - Converge mais rápido que a Iteração de Valor (faz mais de uma iteração na avaliação de política)

Iteração de política truncada

- Três algoritmos de programação dinâmica para encontrar políticas ótimas:
  - Iteração de Valor
    - Baseada na teorema do ponto fixo da equação de Bellman.
    - Duas etapas: atualização de valor e atualização de política.
  - Iteração de Política
    - Mais complexa.
    - Duas etapas: avaliação de política e melhoria de política.
  - Iteração de Política Truncada
    - Unifica os dois algoritmos anteriores como casos extremos.
  - Todos os algoritmos anteriores compartilham uma estrutura comum:
    - Cada iteração tem dois passos:
    - Um para atualizar o valor de estado
    - Outro para atualizar a política
    - Essa ideia é chamada de **Iteração de Política Generalizada**.

Revisão

- **É garantido que o algoritmo de iteração de valor encontre políticas ótimas?**
  - **Sim**. Porque a iteração de valor é exatamente o algoritmo sugerido pelo teorema do mapeamento contrativo para resolver a equação de optimalidade de Bellman. A convergência desse algoritmo é garantida pelo teorema do mapeamento contrativo.
- **Quais etapas estão incluídas no algoritmo de iteração de política?**
  - Cada iteração do algoritmo de iteração de política contém duas etapas: avaliação de política e melhoria de política. Na etapa de **avaliação de política**, o algoritmo busca resolver a equação de Bellman para obter o valor de estado da política atual. Na etapa **de melhoria de política**, o algoritmo busca atualizar a política de forma que a nova política gerada tenha valores de estado maiores.
- **Um outro algoritmo iterativo está embutido no algoritmo de iteração de política?**
  - **Sim**. Na etapa de avaliação de política do algoritmo de iteração de política, é necessário um algoritmo iterativo para resolver a equação de Bellman da política atual.
- **Os valores intermediários gerados pelo algoritmo de iteração de valor são valores de estado?**
  - **Não**. Não há garantia que essas valores satisfaçam a equação de Bellman de nenhuma política.
- **Os valores intermediários gerados pelo algoritmo de iteração de política são valores de estado?**
  - **Sim**. Isso porque esses valores são soluções da equação de Bellman da política atual.

Revisão

- **Qual é a relação entre os algoritmos de iteração de política truncada e iteração de política?**
  - Como o nome sugere, o algoritmo de iteração de política truncada pode ser obtido a partir do algoritmo de iteração de política simplesmente executando um número finito de iterações durante a etapa de avaliação de política.
- **Qual é a relação entre iteração de política truncada e iteração de valor?**
  - A iteração de valor pode ser vista como um caso extremo da iteração de política truncada, em que apenas uma única iteração é executada durante a etapa de avaliação de política.
- **Os valores intermediários gerados pelo algoritmo de iteração de política truncada são valores de estado?**
  - **Não**. Apenas se executarmos um número infinito de iterações na etapa de avaliação de política, poderemos obter os verdadeiros valores de estado. Se executarmos um número finito de iterações, obteremos apenas aproximações dos valores reais de estado.
- **Quantas iterações devemos executar na etapa de avaliação de política do algoritmo de iteração de política truncada?**
  - A diretriz geral é executar poucas iterações, mas não muitas. O uso de poucas iterações na etapa de avaliação de política pode acelerar a taxa de convergência geral, mas executar muitas iterações não acelerará significativamente essa taxa.

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.