Aprendizado por Reforço

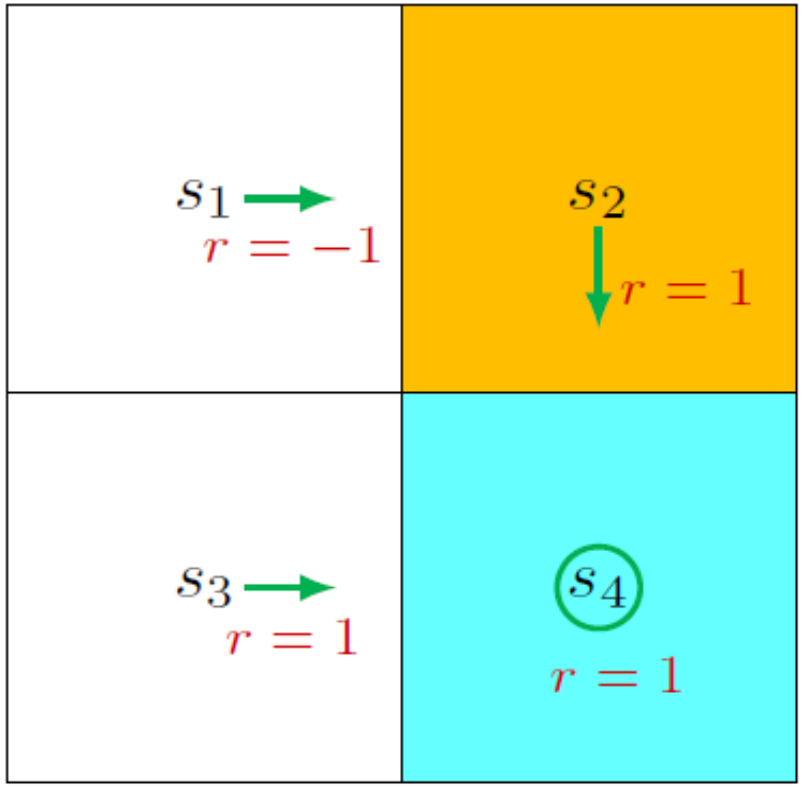
Processos de Decisão de Markov (parte 3)

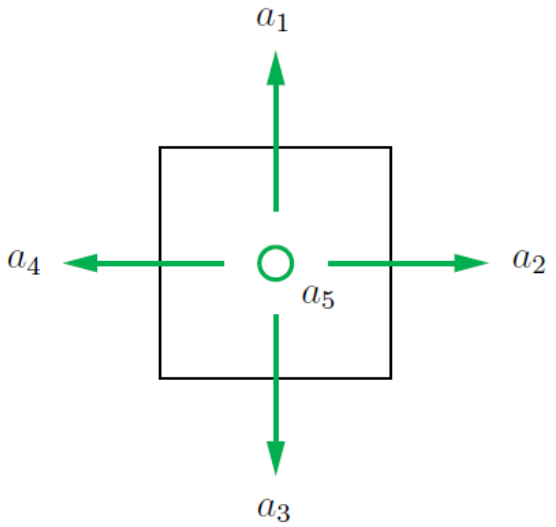
Valores de estado ótimos e Equação de otimalidade de Bellman

- *"Resolver uma tarefa de aprendizado por reforço significa, de forma geral, encontrar uma política que obtenha uma grande recompensa no longo prazo"*. (Sutton e Barto, 2018)
- Conceitos principais:
  - Valores de estado ótimos
  - Políticas ótimas
- Ferramenta-chave:
  - Equação de otimalidade de Bellman
- Cronograma
  - O que nós já vimos: Equação de Bellman (para uma dada política)
  - O que nós veremos hoje: Equação de otimalidade de Bellman (para política ótima)
  - O que nós vermos nas próximas aulas: algoritmos (para resolver a equação de otimalidade de Bellman)

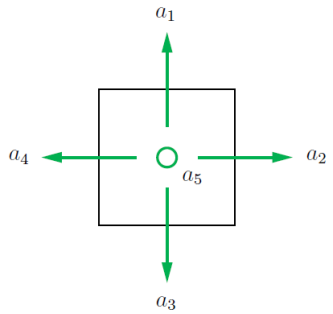
Valores de estado ótimos e Equação de otimalidade de Bellman

- Como melhorar políticas?

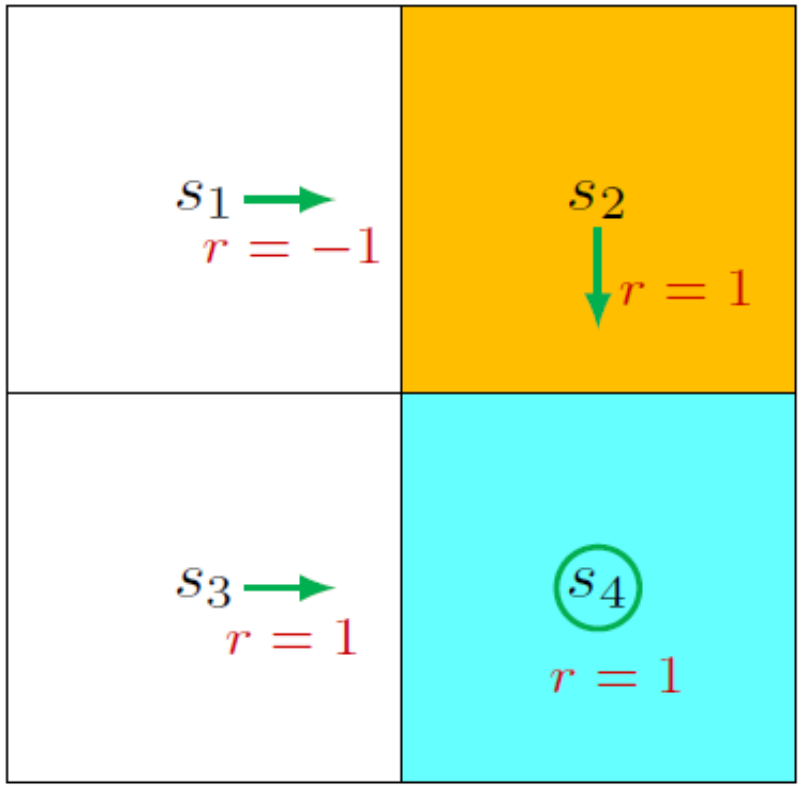




Valores de estado ótimos e Equação de otimalidade de Bellman



- Como melhorar políticas?



Valores de estado (equação de Bellman)

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

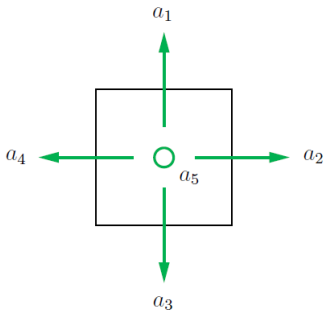
$v_\pi(s_1) = ?$

$v_\pi(s_2) = ?$

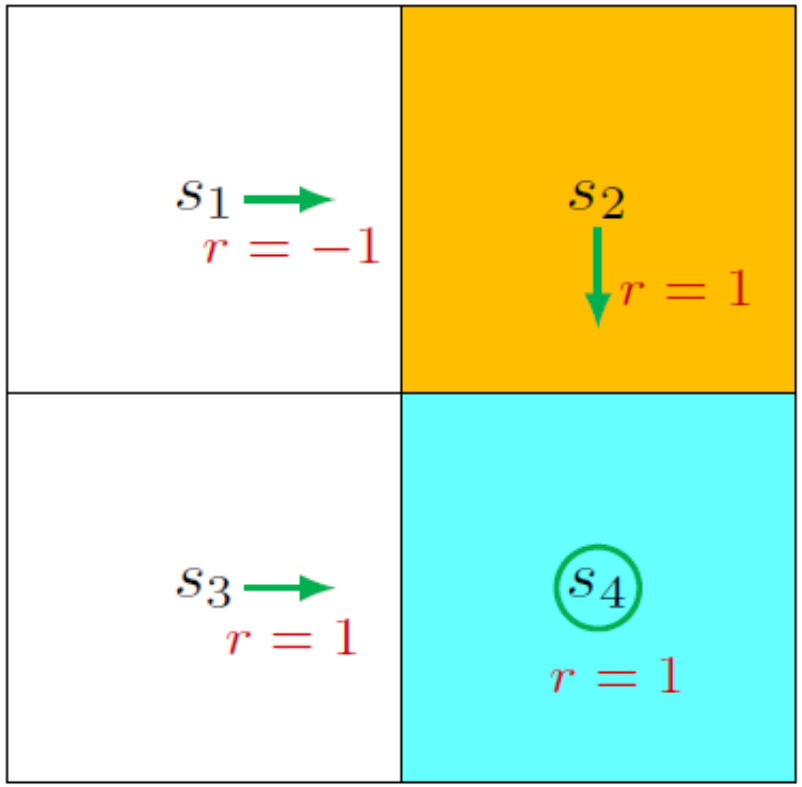
$v_\pi(s_3) = ?$

$v_\pi(s_4) = ?$

Valores de estado ótimos e Equação de otimalidade de Bellman



- Como melhorar políticas?



Valores de estado (equação de Bellman)

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

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

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

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

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

Para $\gamma = 0.9$:

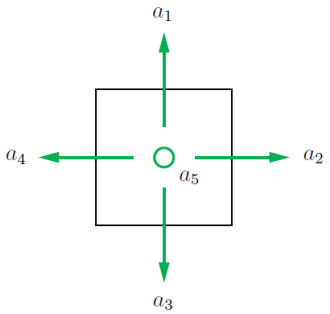
$v_\pi(s_1) = 8$

$v_\pi(s_2) = 10$

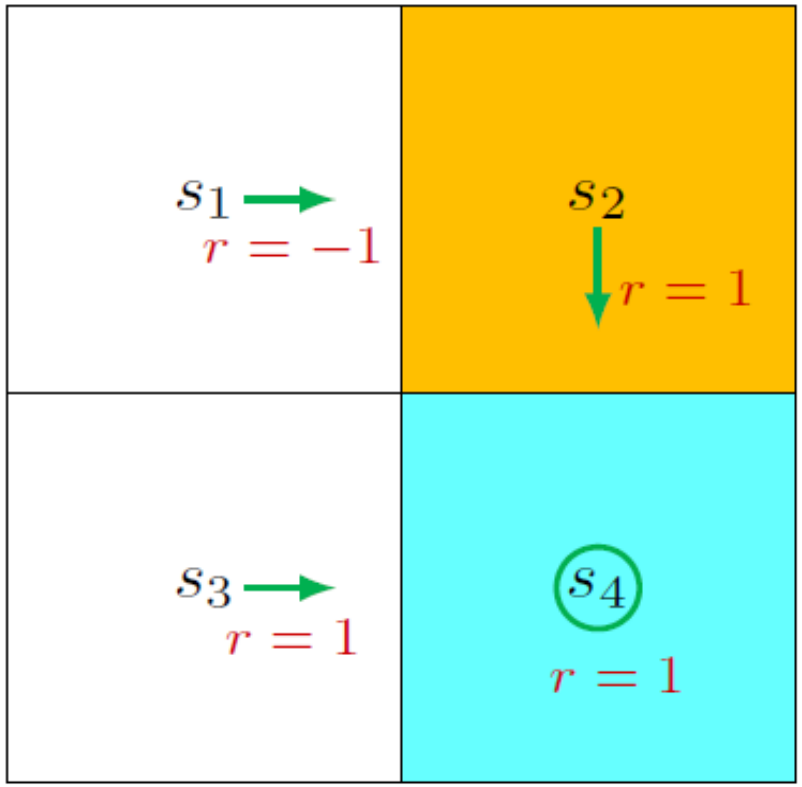
$v_\pi(s_3) = 10$

$v_\pi(s_4) = 10$

Valores de estado ótimos e Equação de otimalidade de Bellman



- Como melhorar políticas?



Valores de ação (equação de Bellman) para o estado s_1

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

$q_\pi(s_1, a_1) = ?$

$q_\pi(s_1, a_2) = ?$

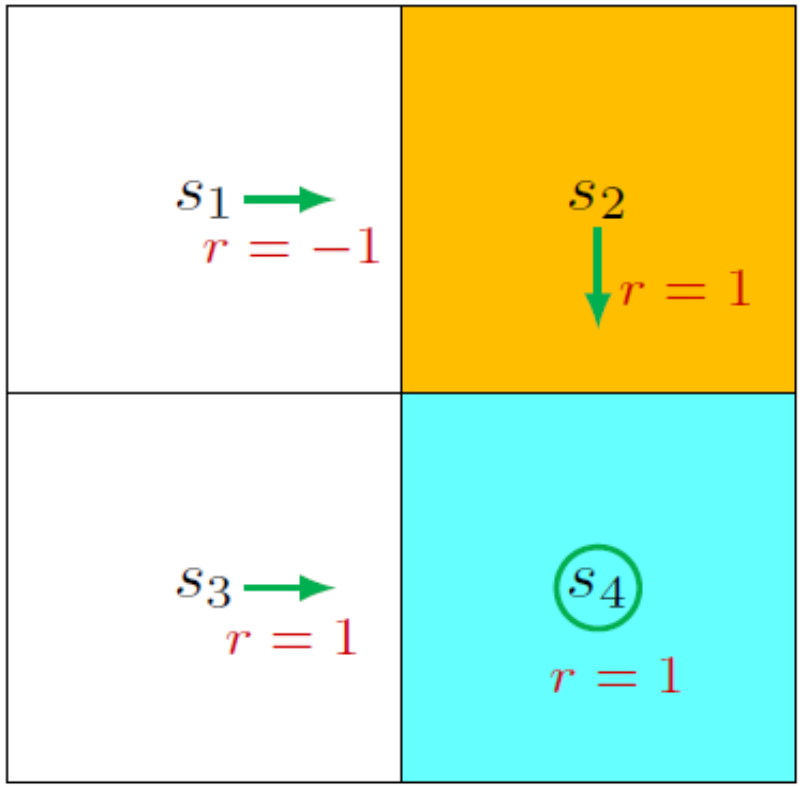
$q_\pi(s_1, a_3) = ?$

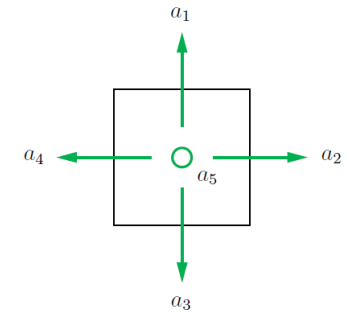
$q_\pi(s_1, a_4) = ?$

$q_\pi(s_1, a_5) = ?$

Valores de estado ótimos e Equação de otimalidade de Bellman

- Como melhorar políticas?





Valores de ação (equação de Bellman) para o estado s_1

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

$$q_\pi(s_1,a_1) = -1 + \gamma v_\pi(s_1) = 6.2$$

$$q_\pi(s_1,a_2) = -1 + \gamma v_\pi(s_2) = 8$$

$$q_\pi(s_1,a_3) = 0 + \gamma v_\pi(s_3) = 9$$

$$q_\pi(s_1,a_4) = -1 + \gamma v_\pi(s_1) = 6.2$$

$$q_\pi(s_1,a_5) = 0 + \gamma v_\pi(s_1) = 7.2$$

Para $\gamma = 0.9$:

$v_\pi(s_1) = 8$

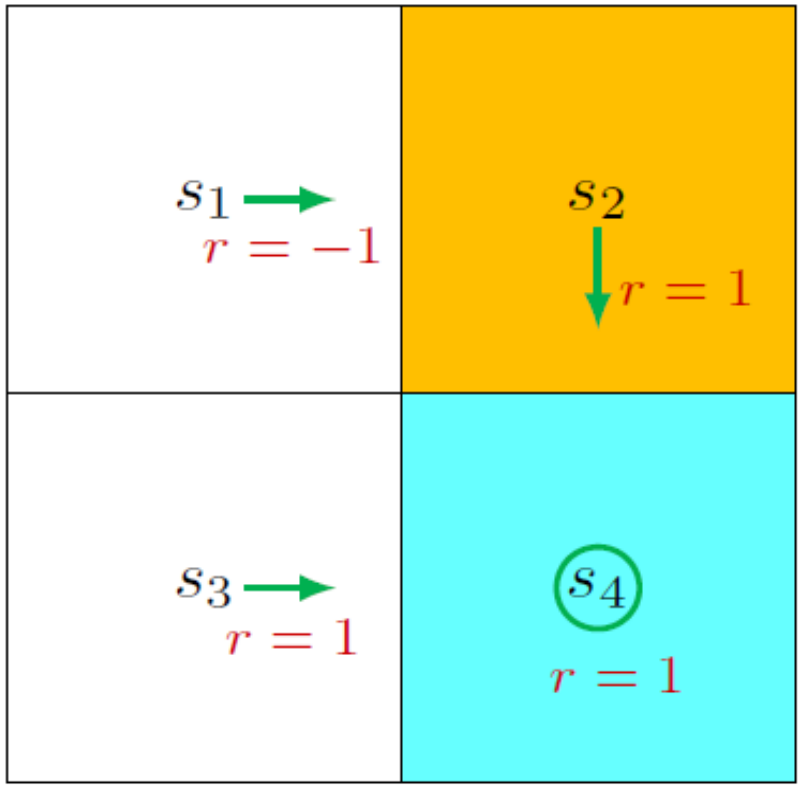
$v_\pi(s_2) = 10$

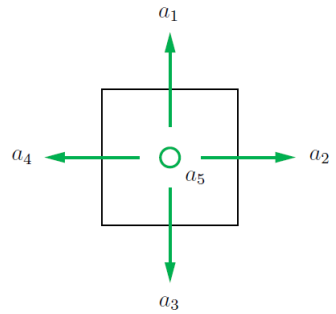
$v_\pi(s_3) = 10$

$v_\pi(s_4) = 10$

Valores de estado ótimos e Equação de otimalidade de Bellman

- Como melhorar políticas?





Valores de ação (equação de Bellman) para o estado s_1

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

$$q_\pi(s_1,a_1) = -1 + \gamma v_\pi(s_1) = 6.2$$

$$q_\pi(s_1,a_2) = -1 + \gamma v_\pi(s_2) = 8$$

$$\boxed{q_\pi(s_1,a_3) = 0 + \gamma v_\pi(s_3) = 9}$$

$$q_\pi(s_1,a_4) = -1 + \gamma v_\pi(s_1) = 6.2$$

$$q_\pi(s_1,a_5) = 0 + \gamma v_\pi(s_1) = 7.2$$

$q_\pi(s_1,a_3) \geq q_\pi(s_1,a_i)$, *para todo* $i \neq 3$

Podemos atualizar a política para executar a ação a_3 no estado s_1!

Atualizar a política para executar a ação com maior valor de ação melhora a política. Essa é a ideia subjacente de muitos algoritmos de aprendizado por reforço.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Valores de estado definem uma ordem parcial entre as políticas:
  - Comparação de políticas: π_1 é melhor do que π_2 se

$$v_{\pi_1}(s) \geq v_{\pi_2}(s), \qquad para\ todo\ s \in \mathcal{S}$$

  - Uma política π^* ótima satisfaz:

$$v_{\pi^*}(s) \geq v_{\pi}(s), \qquad para\ todo\ s \in \mathcal{S}\ e\ política\ \pi$$

  - Valor de estado ótimo

$$v_{\pi^*}(s) = v^* = \max_{\pi} v_{\pi}(s)$$

Valores de estado ótimos e Equação de otimalidade de Bellman

- Questões fundamentais sobre políticas ótimas
  - Existência: a política ótima existe?
  - Unicidade: se existir, a política ótima é única?
  - Estocasticidade: a política ótima é estocástica ou determinística?
  - Algoritmo: como calcular valores ótimos de estado e obter a política ótima?

Valores de estado ótimos e Equação de otimalidade de Bellman

- Equação de Bellman (forma escalar)

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

- Equação de otimalidade de Bellman (forma escalar)

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

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_a \pi(a|s) q(s,a), \qquad s \in \mathcal{S}$$

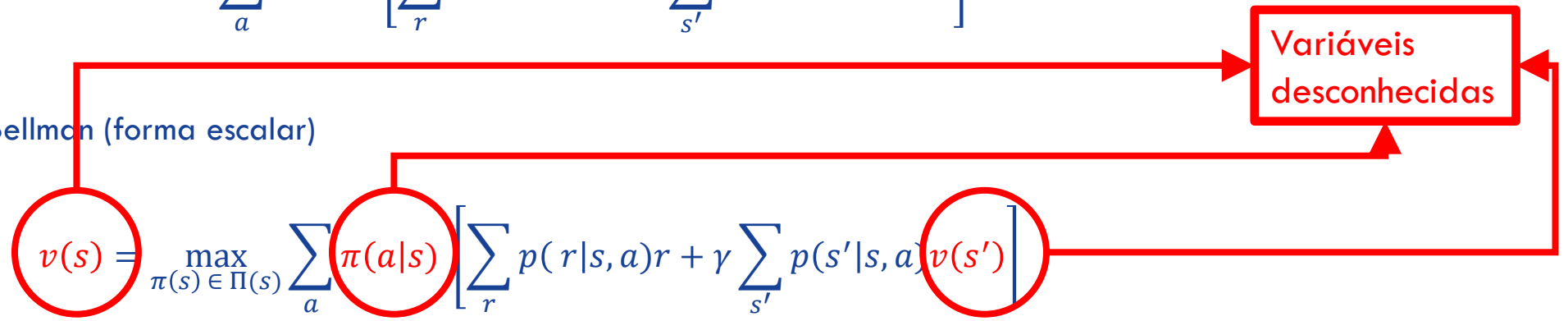
- Onde
  - $q(s,a) = \sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s')$
  - $\pi(s)$: política para o estado s
  - $\Pi(s)$: conjunto de todas as possíveis políticas para s

Valores de estado ótimos e Equação de otimalidade de Bellman

- Equação de Bellman (forma escalar)

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

- Equação de otimalidade de Bellman (forma escalar)



$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_a \pi(a|s) q(s,a), \qquad s \in \mathcal{S}$$

- Onde
  - $q(s,a) = \sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s')$
  - $\pi(s)$: política para o estado s
  - $\Pi(s)$: conjunto de todas as possíveis políticas para s

Valores de estado ótimos e Equação de otimalidade de Bellman

- Equação de otimalidade de Bellman (forma escalar)

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_a \pi(a|s) q(s,a), \qquad s \in \mathcal{S}$$

  - Resolvendo para π

$$Como \sum_a \pi(a|s) = 1, \qquad Então:$$

$$\sum_a \pi(a|s) q(s,a) \leq \sum_a \pi(a|s) \max_{a \in \mathcal{A}} q(s,a) = \max_{a \in \mathcal{A}} q(s,a)$$

$$\sum_a \pi(a|s) q(s,a) = \sum_a \pi(a|s) \max_{a \in \mathcal{A}} q(s,a) \qquad se \quad \pi(a|s) = \begin{cases} 1, & a = a^* \\ 0, & a \neq a^* \end{cases}, \qquad onde\ a^* = \operatorname*{argmax}_a q(s,a)$$

  - A política ótima é aquela que seleciona a ação que tem o maior valor de $q(s,a)$.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Questões fundamentais a equação de otimalidade de Bellman
  - Existência: Essa equação tem solução?
  - Unicidade: Se existir, a solução é única?
  - Algoritmo: Como resolver essa equação?
  - Otimalidade: Qual a relação da solução com políticas ótimas?

Valores de estado ótimos e Equação de otimalidade de Bellman

- Equação de otimalidade de Bellman na forma matricial

$$v = f(v) \triangleq \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v)$$

Onde

$$v \in \mathbb{R}^{|\mathcal{S}|}$$

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

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

- O operador $\max_{\pi \in \Pi}$ é aplicado "elemento a elemento"
- Resolver $v = f(v)$ para encontrar o ponto fixo que representa os valores de estado ótimos

Valores de estado ótimos e Equação de otimalidade de Bellman

- Teorema do ponto fixo
  - O que é um ponto fixo?
    - Seja uma função $f(x)$, onde $x \in \mathbb{R}^d$ e $f: \mathbb{R}^d \to \mathbb{R}^d$. Um ponto x^* é chamado um ponto fixo se

$$\boxed{f(x^*) = x^*}$$

    - Isto é, x^* é mapeado para si mesmo.
  - O que é uma função contrativa?
    - A função f é uma função contrativa se existir $\gamma \in (0,1)$ tal que

$$\boxed{\|f(x_1) - f(x_2)\| \leq \gamma \|x_1 - x_2\|}$$

para qualquer $x_1, x_2 \in \mathbb{R}^d$. ($\|\cdot\|$ representa uma norma vetorial ou matricial)

Valores de estado ótimos e Equação de otimalidade de Bellman

- Teorema do ponto fixo
  - Para qualquer equação da forma $x = f(x)$, em que x e $f(x)$ são vetores reais, se f é um mapeamento de contração, então as seguintes propriedades valem:
    - Existência: Existe um ponto fixo x^* que satisfaz $f(x^*) = x^*$.
    - Unicidade: Esse ponto fixo x^* é único.
    - Algoritmo: o processo iterativo a seguir pode ser utilizado para encontrar a solução

Solução única

$$x_{k+1} = f(x_k)$$

Convergência

para $k = 0, 1, 2, \ldots$. Então, $x_k \to x^*$ à medida que $k \to \infty$ para qualquer valor inicial x_0.

Valores de estado ótimos e Equação de otimalidade de Bellman

- No contexto da equação de otimalidade de Bellman, $f(v)$ é um mapeamento de contração. Para qualquer $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$:

$$\boxed{\| f(v_1) - f(v_2) \|_\infty \leq \gamma \| v_1 - v_2 \|_\infty}$$

$\gamma \in (0,1)$: taxa de desconto

$\|\cdot\|_\infty$: norma ℓ_∞ (valor absoluto máximo dos elementos do vetor)

- Podemos aplicar diretamente o teorema do ponto fixo para analisar a equação de otimalidade de Bellman.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Resolvendo a equação de otimalidade de Bellman para Encontrar v^* e π^*.
  - Se v^*é uma solução para a equação de otimalidade de Bellman, então v^* satisfaz

$$v^* = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v^*)$$

  - v^* é um ponto fixo de $f(v)$ $(v^* = f(v^*))$, então:
    - Existência de v^* : a solução para a equação de otimalidade de Bellman sempre existe.
    - Unicidade de v^*: a solução v^* para a equação de otimalidade de Bellman é sempre única.
    - Algoritmo (**iteração de valor**): O valor de v^* pode ser encontrado por um algoritmo iterativo

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

O valor de v_k converge para v^* quando $k \to \infty$, dado qualquer palpite inicial v_0.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Resolvendo a equação de otimalidade de Bellman para Encontrar v^* e π^*.
  - Após obter v^* Podemos extrair a política ótima π^*:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}}(r_\pi + \gamma P_\pi v^*)$$

  - Relação entre a equação de otimalidade de Bellman e a equação de Bellman:

$$v^* = v_{\pi^*} = r_{\pi^*} + \gamma P_{\pi^*} v^*$$

    - A equação de otimalidade de Bellman é um caso especial da equação de Bellman cuja política é π^*.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Otimalidade de v^* e π^*
  - A solução v^* é o valor de estado ótimo, e π^* é uma política ótima. Isto é, para qualquer política π, vale:

$$v^* = v_{\pi^*} \geq v_\pi$$

em que v_π é o valor de estado da política π, e $\geq$ denota comparação elemento a elemento.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Política gulosa ótima
  - Para cada $s \in \mathcal{S}$, a política gulosa determinística

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

é uma política ótima para resolver a equação de otimalidade de Bellman. Onde,

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

Valor de ação ótimo:

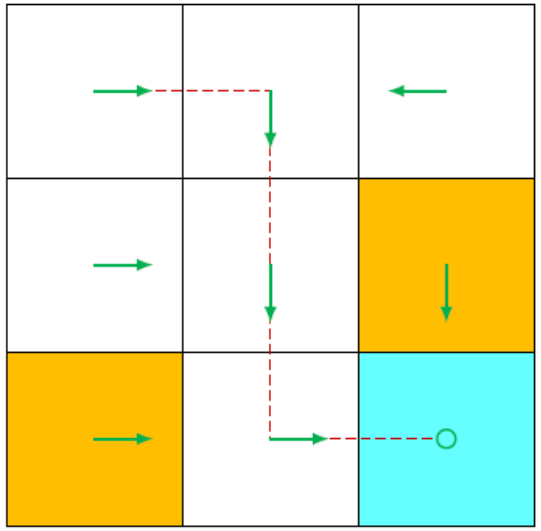
$$q^*(s,a) = \sum_{r \in \mathcal{R}} p(r|s,a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)v^*(s')$$

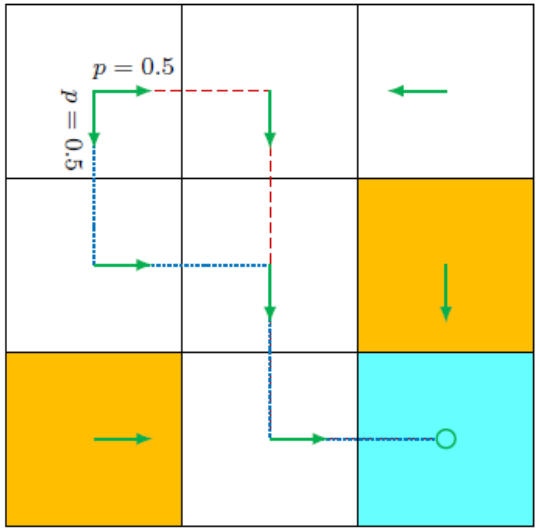
  - Busca as ações com maior $q^*(s,a)$.

Valores de estado ótimos e Equação de otimalidade de Bellman

- Política ótima π^*
  - Unicidade de π^*: Embora v^*seja único, π^* pode não ser.
  - Estocasticidade de π^*: Uma política ótima pode ser estocástica ou determinística. No entanto, é certo que sempre existe ao menos uma política ótima determinística.

Política ótima determinística





Política ótima estocástica

Valores de estado ótimos e Equação de otimalidade de Bellman

- Equação de otimalidade de Bellman

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_{a} \pi(a|s) \left[\sum_{r} p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right]$$

- Quais parâmetros determinam v^* e π^*?

Valores de estado ótimos e Equação de otimalidade de Bellman

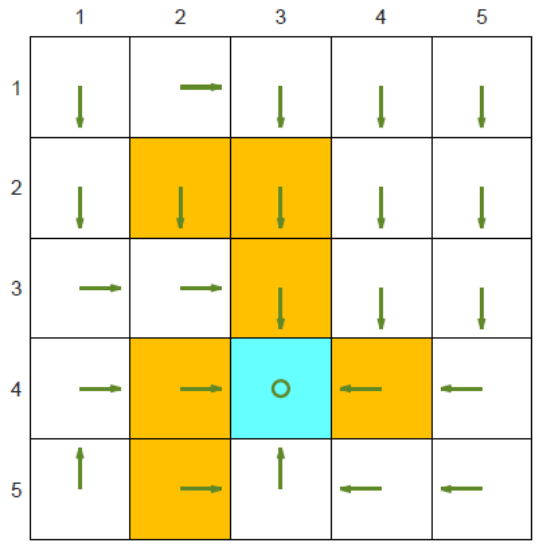
- Equação de otimalidade de Bellman

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_{a} \pi(a|s) \left[\sum_{r} p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right]$$

- Quais parâmetros determinam v^* e π^*?
  - Recompensas imediata (r)
  - Taxa de desconto (γ)
  - Modelo do ambiente $(p(r|s,a), p(s'|s,a))$

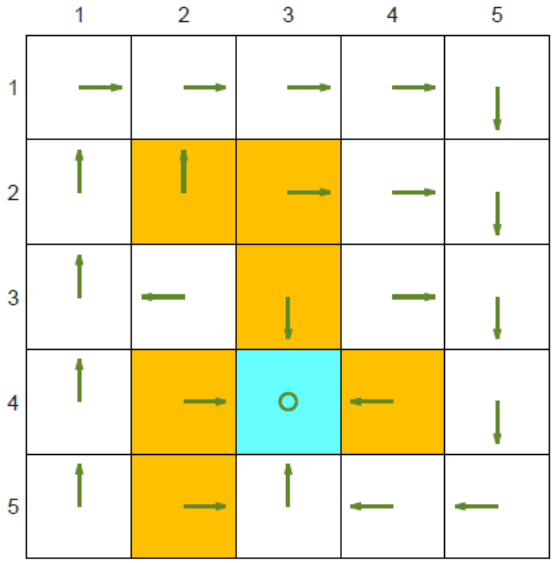
Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da taxa de desconto (γ).



	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

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

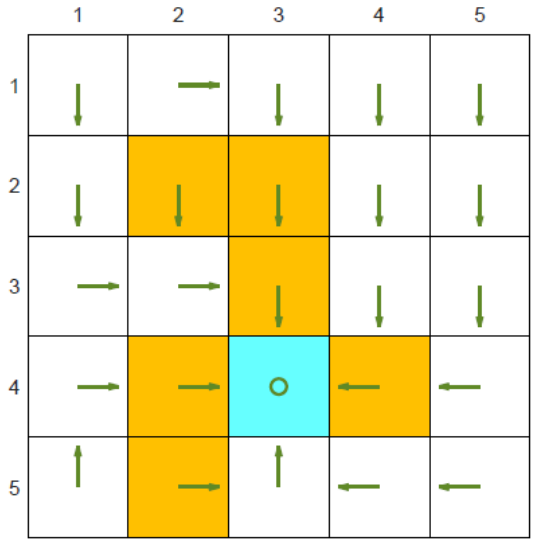


	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	2.0	0.1	0.1
4	0.0	2.0	2.0	2.0	0.2
5	0.0	1.0	2.0	1.0	0.5

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

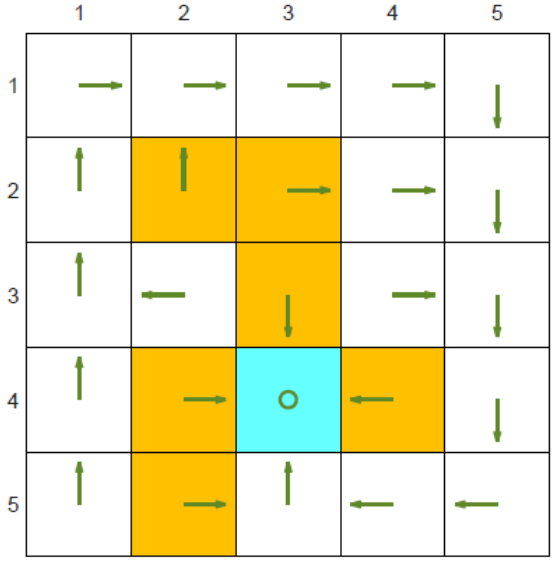
Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da taxa de desconto (γ).



	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

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

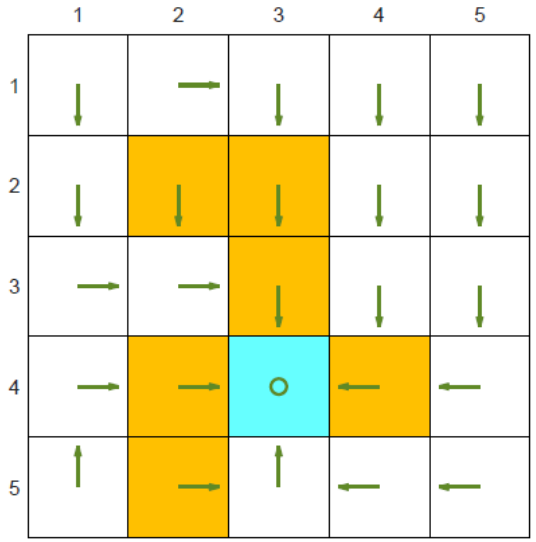


	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	2.0	0.1	0.1
4	0.0	2.0	2.0	2.0	0.2
5	0.0	1.0	2.0	1.0	0.5

$$r_{boundary} = r_{forbidden} = -1$$
$$r_{target} = 1$$
$$\gamma = 0.5$$

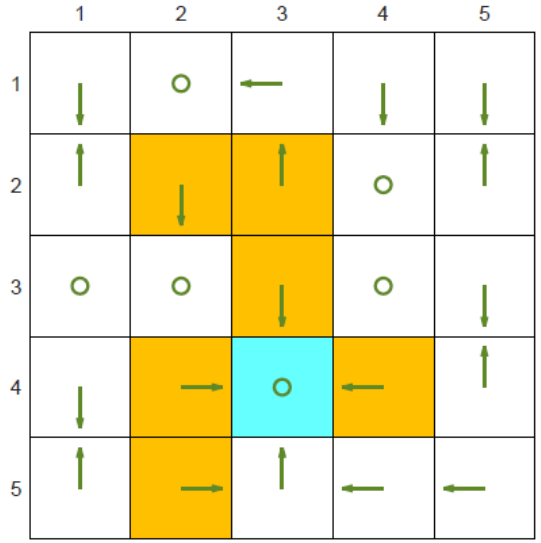
Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da taxa de desconto (γ).



	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

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



	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	1.0	1.0	1.0	0.0
5	0.0	0.0	1.0	0.0	0.0

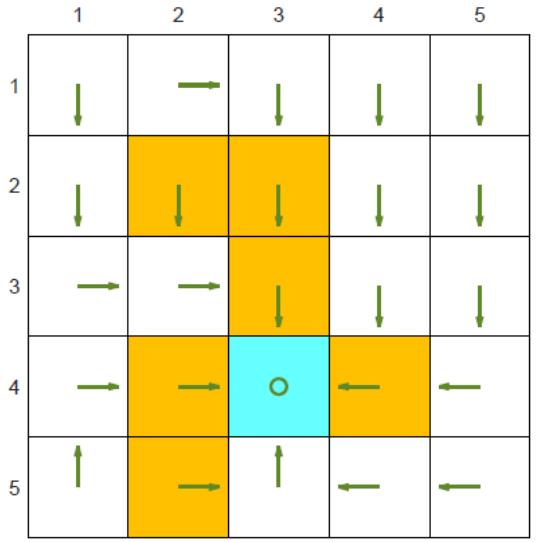
$$r_{boundary} = r_{forbidden} = -1$$
$$r_{target} = 1$$
$$\gamma = 0.0$$

Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da taxa de desconto (γ).
  - Alterar γ pode tornar a política mais "corajosa" (se γ for alto) ou mais "cautelosa" (se γ baixo).
  - Distribuição espacial dos valores de estado
    - Os estados mais próximos do alvo têm valores de estado mais altos, enquanto os mais distantes têm valores mais baixos.
    - Pode ser explicado pela taxa de desconto: se um estado precisa de mais passos para chegar ao alvo, seu valor de estado diminui devido ao desconto.

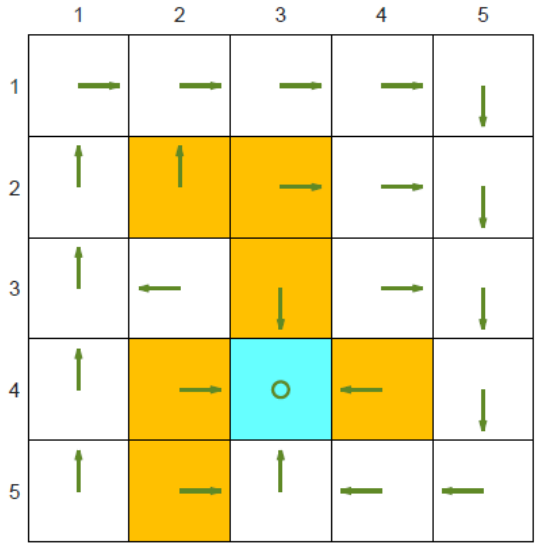
Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da recompensas imediata (r).



	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

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

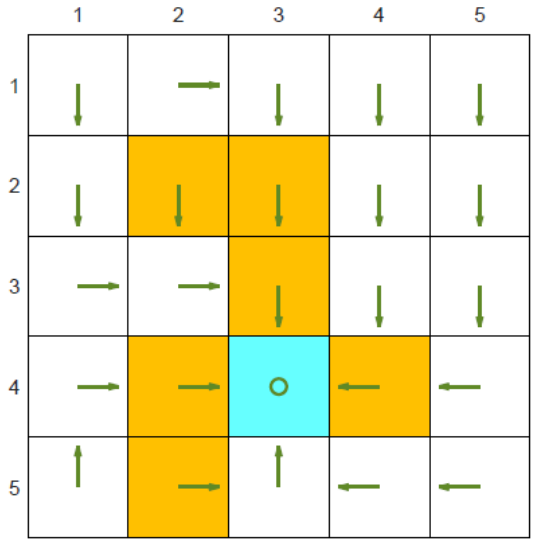


	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

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

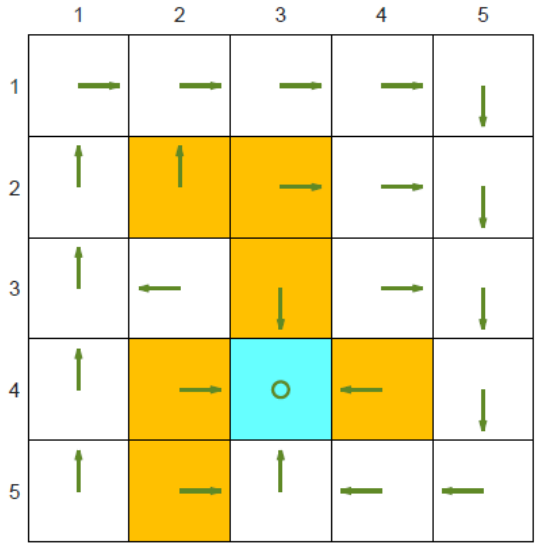
Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da recompensas imediata (r).



	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

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



	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

$$r_{boundary} = -1$$
$$r_{forbidden} = -10$$
$$r_{target} = 1$$
$$\gamma = 0.9$$

Valores de estado ótimos e Equação de otimalidade de Bellman

- Fatores que influenciam políticas ótima: impacto da recompensas imediata (r).
  - Punir mais fortemente (recompensa negativa maior) desencoraja certas ações.

Mas cuidado!!!

  - Mudar as recompensas nem sempre leva a políticas ótimas diferentes.
  - As políticas ótimas são invariantes a transformações afins das recompensas:
    - Considere um processo de decisão de Markov cujo valor de estado ótimo v^* satisfaz

$$v^* = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v^*)$$

Se todas as recompensa $r \in \mathbb{R}$ forem modificadas por uma transformação afim para $\alpha r + \beta$, onde $\alpha, \beta \in R$ e $\alpha > 0$, então o valor de estado ótimo correspondente v' também será uma transformação afim de v^*:

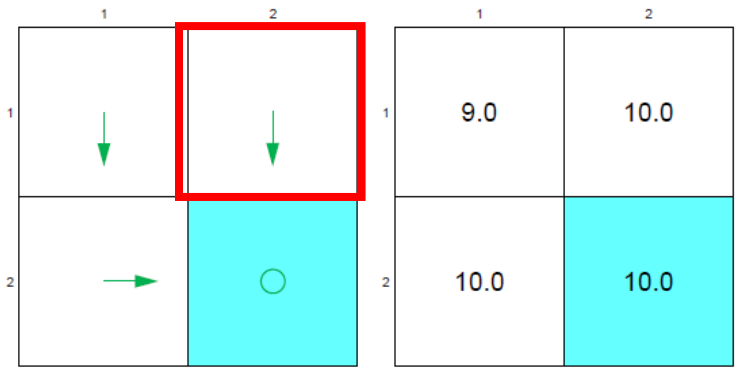
$$v' = \alpha v^* + \frac{\beta}{1-\gamma}\mathbf{1}$$

onde $\gamma \in (0,1)$ é a taxa de desconto e $\mathbf{1} = [1, \ldots, 1]^\top$.

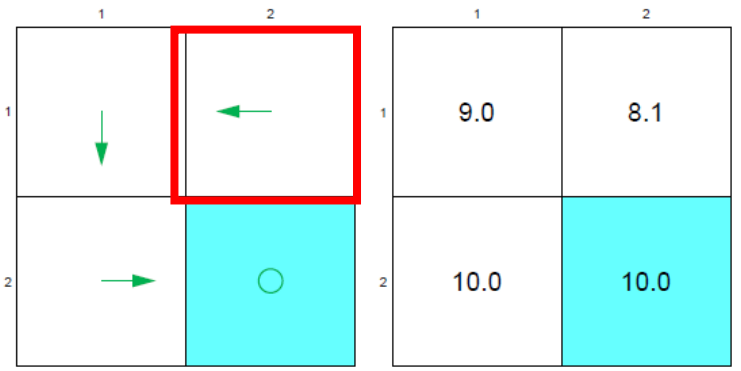
- Consequentemente, a política ótima derivada de v' é invariável em relação à transformação afim dos valores de recompensa!

Valores de estado ótimos e Equação de otimalidade de Bellman

- Acelerando o Alcance ao Objetivo
  - $r_{other} = 0$ não é uma punição. Será que a política ótima faria desvios desnecessários antes de chegar ao alvo?



Política ótima



Política não ótima

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

Valores de estado ótimos e Equação de otimalidade de Bellman

- Acelerando o Alcance ao Objetivo
  - Considerando apenas recompensas imediatas, o desvio não teria problema, pois não envolve recompensas negativas imediatas.
  - Considerando o retorno descontado, o desvio passa a importar.
    - Política ótima:

$$retorno = 1 + \gamma\, 1 + \gamma^2\, 1 + \cdots = \frac{1}{1-\gamma} = 10, \qquad (\gamma = 0.9)$$

    - Política não ótima

$$retorno = 0 + \gamma\, 0 + \gamma^2\, 1 + \gamma^3\, 1 + \cdots = \frac{\gamma^2}{1-\gamma} = 8.1, \qquad (\gamma = 0.9)$$

  - Mesmo com recompensas "0" para passos intermediários, a taxa de desconto (γ) faz a política evitar caminhos longos (quanto menor a trajetória, maior o retorno).

Valores de estado ótimos e Equação de otimalidade de Bellman

- Conceitos:
  - valores de estado ótimos
  - políticas ótimas: seus valores de estado são maiores ou iguais aos valores de qualquer outra política
- Equação de otimalidade de Bellman
  - Equação não linear com propriedade de contração (existência e unicidade de v^*).
  - Soluções correspondem tanto aos valores de estado ótimos quanto às políticas ótimas.
  - Base para algoritmos de aprendizado por reforço.

Referências

- Shiyu Zhao. Mathematical Foundations of Reinforcement Learning. Springer Singapore, 2025. [capítulo 3]
  - 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.