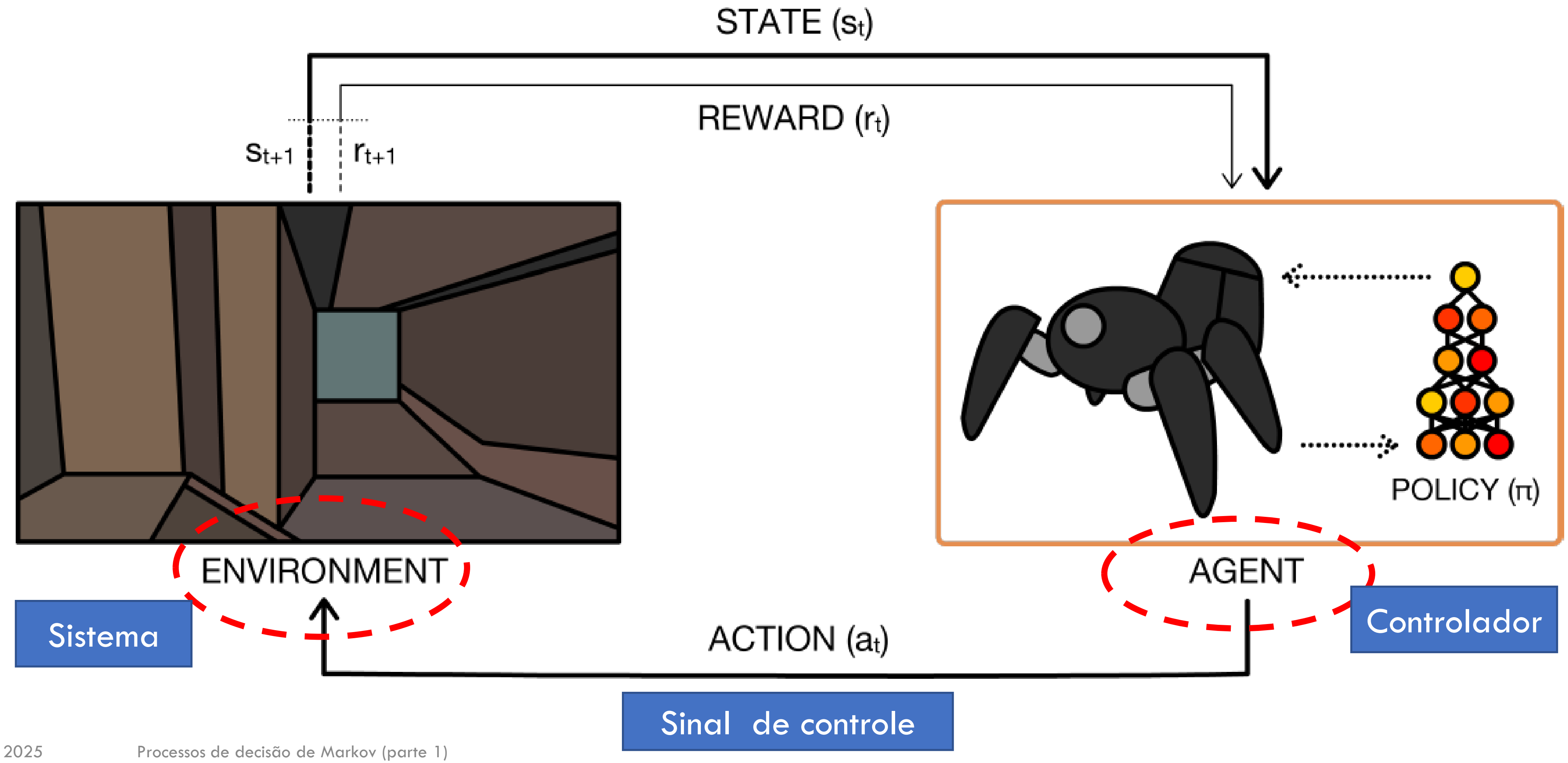
Aprendizado por Reforço

Processos de Decisão de Markov
(parte 1)

Revisão da aula passada...

- Paradigma de aprendizado por <u>interação:</u> Agente ⟺ Ambiente

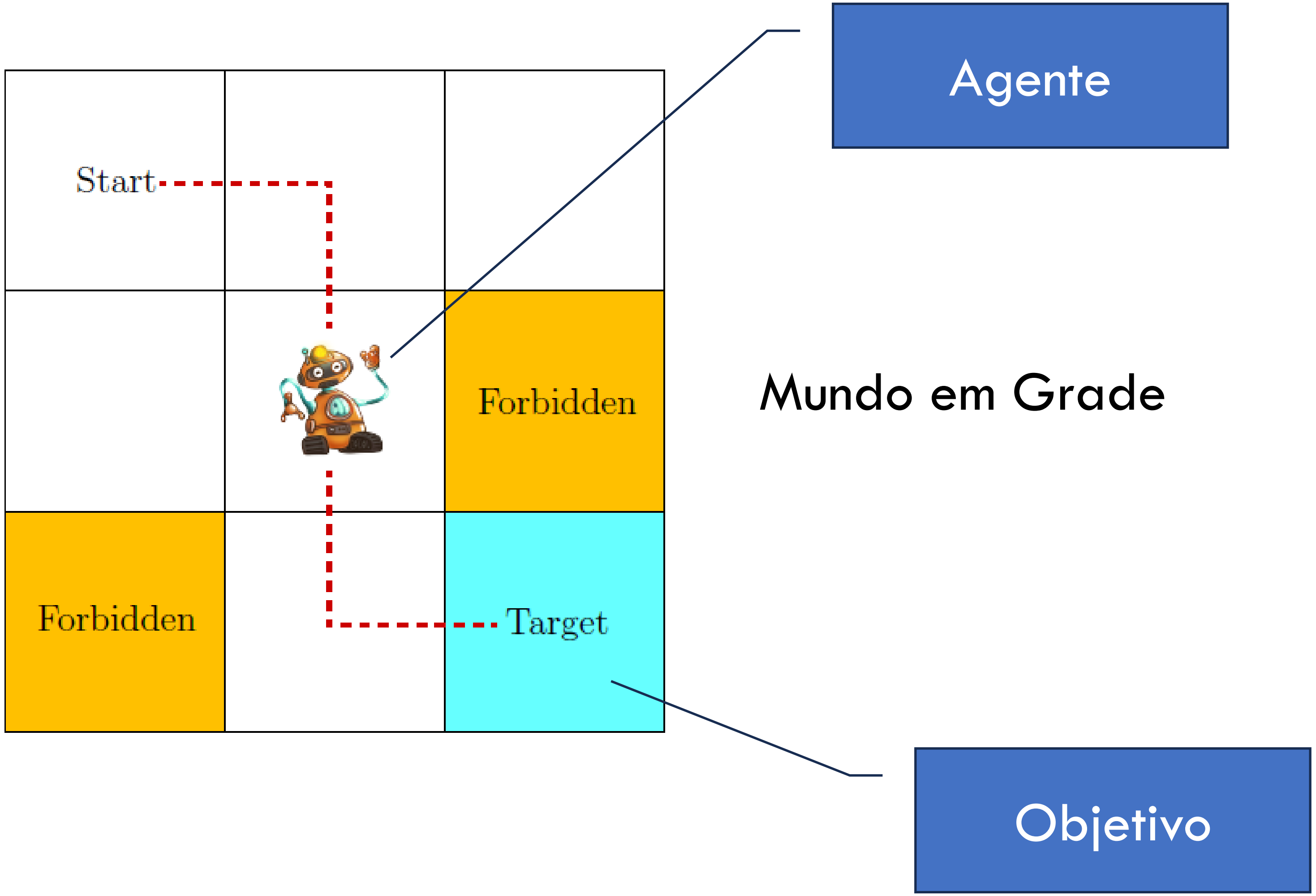


K. Arulkumaran et al., "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

Revisão da aula passada...

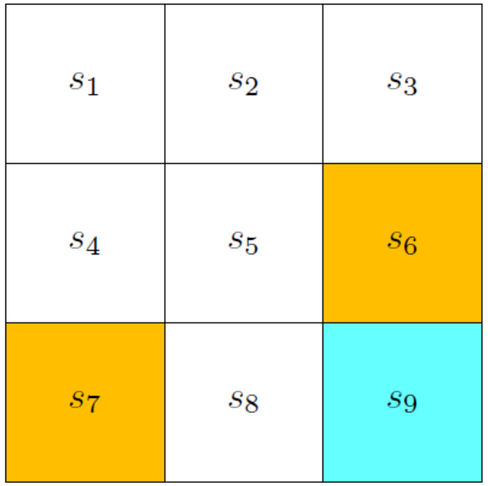
Tarefa não-trivial quando o agente não tem nenhuma informação a priori sobre o ambiente!

Queremos encontrar uma política para alcançar o alvo.

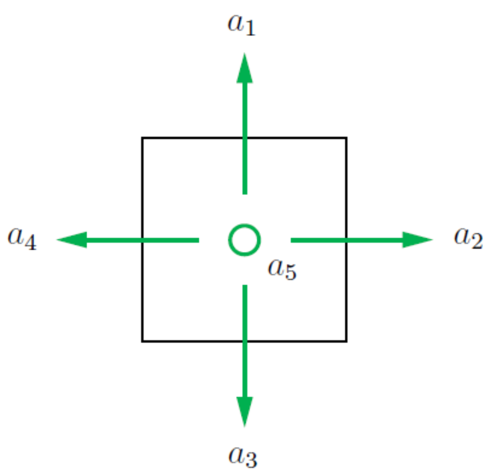


Mundo em Grade

Revisão da aula passada...



(a) States



(b) Actions

Espaço de estados: $\mathcal{S} = \{s_1, \dots, s_9\}$
Espaço de ações: $\mathcal{A} = \{a_1, \dots, a_5\}$
(Pode ser uma função do estado $\mathcal{A}(s_i)$)

Revisão da aula passada...

Transição de estados

	a_1 (upward)	a_2 (rightward)	a_3 (downward)	a_4 (leftward)	a_5 (still)
s_1	s_1	s_2	s_4	s_1	s_1
s_2	s_2	s_3	s_5	s_1	s_2
s_3	s_3	s_3	s_6	s_2	s_3
s_4	s_1	s_5	s_7	s_4	s_4
s_5	s_2	s_6	s_8	s_4	s_5
s_6	s_3	s_6	s_9	s_5	s_6
s_7	s_4	s_8	s_7	s_7	s_7
s_8	s_5	s_9	s_8	s_7	s_8
s_9	s_6	s_9	s_9	s_8	s_9

$$p(s_1|s_1,a_2) = 0, \quad p(s_2|s_1,a_2) = 1, \quad p(s_3|s_1,a_2) = 0,$$
$$p(s_4|s_1,a_2) = 0, \quad p(s_5|s_1,a_2) = 0$$

Recompensa

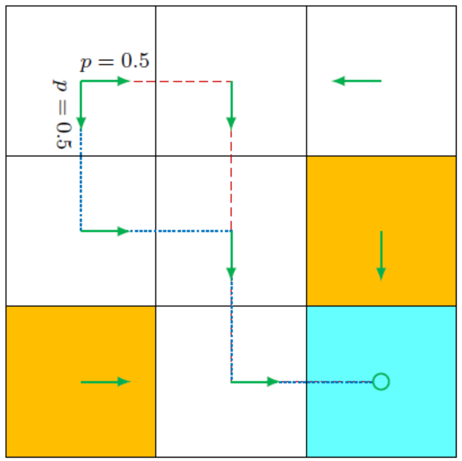
	a_1 (upward)	a_2 (rightward)	a_3 (downward)	a_4 (leftward)	a_5 (still)
s_1	r_{boundary}	0	0	r_{boundary}	0
s_2	r_{boundary}	0	0	0	0
s_3	r_{boundary}	r_{boundary}	$r_{\text{forbidden}}$	0	0
s_4	0	0	$r_{\text{forbidden}}$	r_{boundary}	0
s_5	0	$r_{\text{forbidden}}$	0	0	0
s_6	0	r_{boundary}	r_{target}	0	$r_{\text{forbidden}}$
s_7	0	0	r_{boundary}	r_{boundary}	$r_{\text{forbidden}}$
s_8	0	r_{target}	r_{boundary}	$r_{\text{forbidden}}$	0
s_9	$r_{\text{forbidden}}$	r_{boundary}	r_{boundary}	0	r_{target}

$$p(r = -1|s_1,a_1) = 1, \quad p(r \neq -1|s_1,a_1) = 0$$

Política π

	a_1 (upward)	a_2 (rightward)	a_3 (downward)	a_4 (leftward)	a_5 (still)
s_1	0	0.5	0.5	0	0
s_2	0	0	1	0	0
s_3	0	0	0	1	0
s_4	0	1	0	0	0
s_5	0	0	1	0	0
s_6	0	0	1	0	0
s_7	0	1	0	0	0
s_8	0	1	0	0	0
s_9	0	0	0	0	1

$$\pi(a_1|s_1) = 0$$
$$\pi(a_2|s_1) = 0.5$$
$$\pi(a_3|s_1) = 0.5$$
$$\pi(a_4|s_1) = 0$$
$$\pi(a_5|s_1) = 0$$



Revisão da aula passada...

- Trajetórias, retornos e episódios

  - As ações tomadas devem ser determinadas pelo retorno (recompensa total) ao invés da recompensa imediata.

  - Trajetórias infinitas:

$$s_1 \underset{r=0}{\overset{a_2}{\rightarrow}} s_2 \underset{r=0}{\overset{a_3}{\rightarrow}} s_5 \underset{r=0}{\overset{a_3}{\rightarrow}} s_8 \underset{r=1}{\overset{a_2}{\rightarrow}} s_9 \underset{r=1}{\overset{a_5}{\rightarrow}} s_9 \underset{r=1}{\overset{a_5}{\rightarrow}} s_9 \cdots$$

  - Problema:

$$retorno \ = 0 + 0 + 0 + 1 + 1 + 1 + \cdots = \infty$$

  - Solução: retorno descontado (introdução de um fator de desconto $\gamma \in (0, 1)$)

$$retorno\ descontado = 0 + \gamma 0 + \gamma^2 0 + \gamma^3 1 + \gamma^4 1 + \gamma^5 1 + \cdots$$

  - O fator de desconto é utilizado para ponderar recompensas no tempo

Continuação da aula passada...

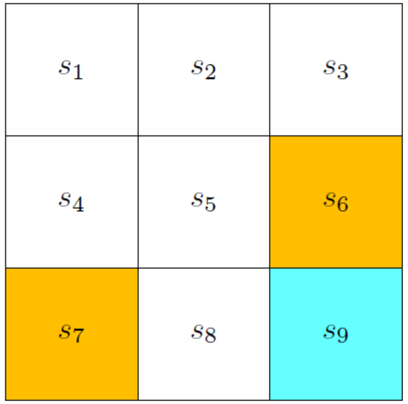
- Episódio (ou tentativa)
  - Um episódio é uma trajetória que termina quando o agente, ao interagir com o ambiente, alcança um **estado terminal**.
    - Ambientes ou política estocásticos: episódios podem ser diferentes mesmo partindo do mesmo estado.
    - Ambientes e políticas determinísticos: iniciar no mesmo estado sempre resulta no mesmo episódio.

Continuação da aula passada...

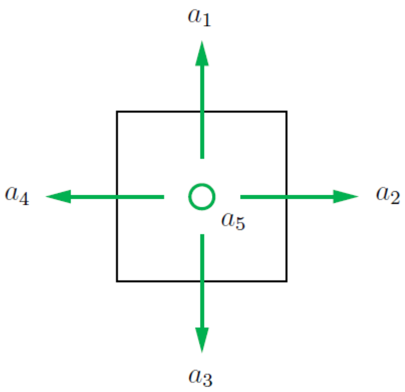
- Tarefas Episódicas vs. Contínuas
  - Tarefas episódicas: possuem trajetórias finitas, terminando em um estado terminal.
  - Tarefas contínuas: não possuem estados terminais e a interação com o ambiente não tem fim.
  - **Como converter tarefas episódicas em tarefas contínuas?**

Continuação da aula passada...

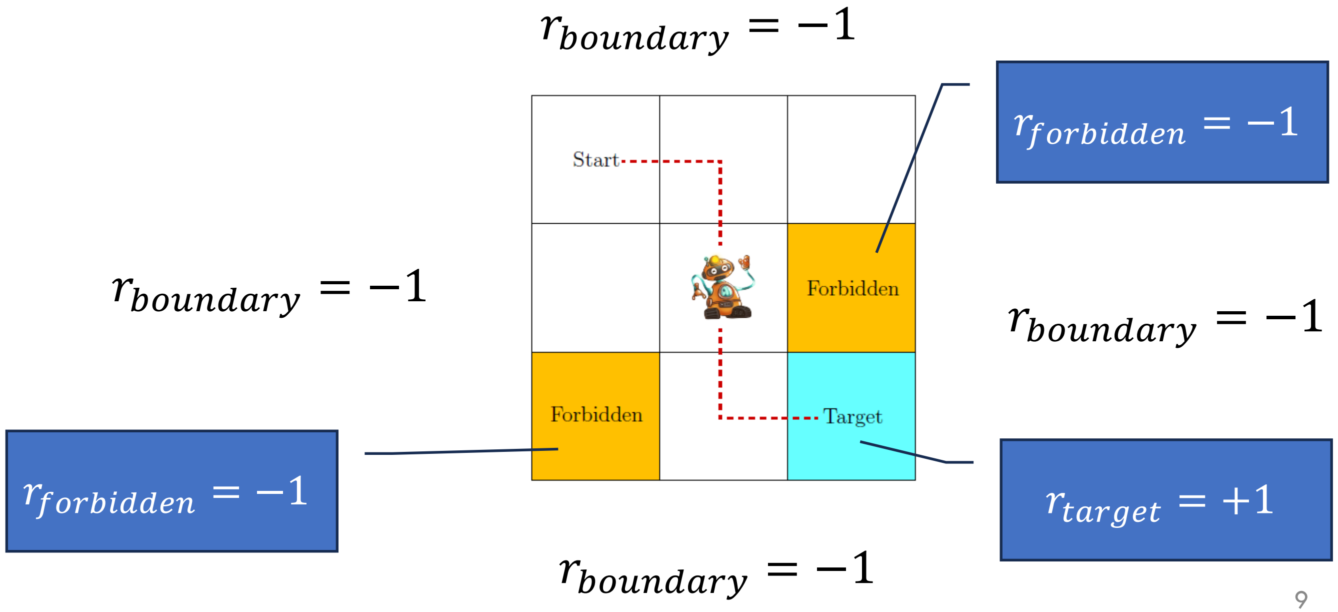
- Conversão de tarefas episódicas em tarefas contínuas: temos que definir regras para o comportamento do agente após alcançar um estado terminal.
  - Modelagem 1 de estados terminais
    - estado terminal como estado absorvente: o agente permanece nesse estado indefinidamente.
    - $\mathcal{A}(s_9) = \{a_5\}$ ou $\mathcal{A}(s_9) = \{a_1, \cdots, a_5\}$ com $p(s_9|s_9, a_i) = 1$ para $i = 1, \cdots 5$.



(a) States

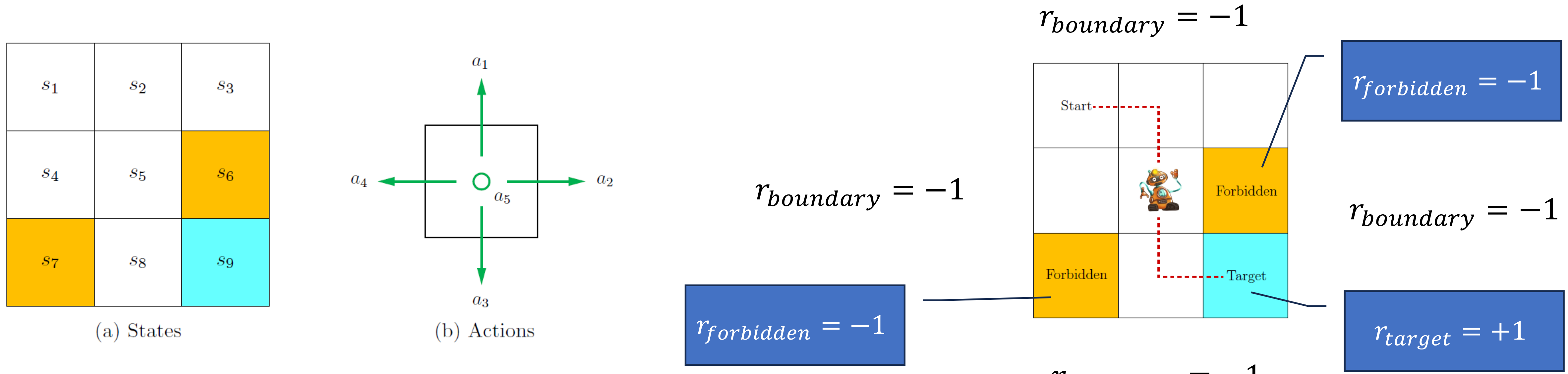


(b) Actions



Continuação da aula passada...

- Modelagem 2 de estados terminais
  - Estado terminal como estado normal: o agente pode continuar interagindo e retornar ao estado terminal.
  - Se receber uma recompensa positiva por alcançar esse estado, o agente pode aprender a permanecer nele para maximizar os ganhos.
  - Para evitar crescimento infinito da recompensa, um fator de desconto ($0 \leq \gamma < 1$) deve ser aplicado ao cálculo do retorno.



Processos de Decisão de Markov

- Processo de Decisão de Markov: é um modelo matemático utilizado para descrever sistemas dinâmicos estocásticos.
- Componentes de um PDM
  - Conjuntos
    - Conjunto de estados ($\mathcal{S}$): conjunto de todos os estados possíveis.
    - Conjunto de ações ($\mathcal{A}(s), s \in \mathcal{S}$): conjunto de ações disponíveis em cada estado.
    - Conjunto de recompensas ($\mathcal{R}(s, a)$): conjunto dos valores de recompensa associados a pares estado-ação.
    - PDMs finitos: o número de estados, ações e recompensas é finito.

Processos de Decisão de Markov

- Componentes de um PDM
  - Modelo
    - Probabilidade de transição ($p(s'|s,a)$): define a probabilidade de transição do estado s' para o estado s após executar uma ação a.

$$p(s'|s,a) \triangleq \Pr\{S_t = s' | S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r | s, a)$$

$$\sum_{s' \in \mathcal{S}} p(s'|s,a) = 1 \text{ para qualquer par } (s,a)$$

    - Probabilidade de recompensa ($p(r|s,a)$): descreve a probabilidade de obter determinada recompensa ao executar uma ação a em um estado s.

$$\sum_{r \in \mathcal{R}(s,a)} p(r|s,a) = 1 \text{para qualquer par } (s,a)$$

      - Recompensa esperada (par ação-estado)

$$r(s,a) \triangleq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r | s, a)$$

      - Recompensa esperada (par ação-estado-próximo estado)

$$r(s,a,s') \triangleq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r | s, a)}{p(s'|s,a)}$$

Processos de Decisão de Markov

- Componentes de um PDM
  - Modelo
    - Modelo ou dinâmica:

$$p(s'|s,a) \; e \; p(r|s,a) \; para \; todo \; par \; (s,a)$$

$$p(s',r|s,a) \triangleq \Pr\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\}$$

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s',r|s,a) = 1, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}(s)$$

    - Tipos de modelos
      - Estacionário: o ambiente não muda com o tempo.
      - Não-estacionário: o ambiente pode mudar ao longo do tempo.

Processos de Decisão de Markov

- Componentes de um PDM
  - Política ($\pi(a|s)$)
    - define a probabilidade de escolher uma ação a em um determinado estado s.
    - $\sum_{a \in \mathcal{A}(s)} \pi(a|s) = 1$ para qualquer estado $s \in \mathcal{S}$.
  - Propriedade de Markov
    - O próximo estado e a próxima recompensa dependem **apenas** do estado e ação atuais.

$$p(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_0, a_0) = p(s_{t+1}|s_t, a_t)$$

$$p(r_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_0, a_0) = p(r_{t+1}|s_t, a_t)$$

Processos de Decisão de Markov

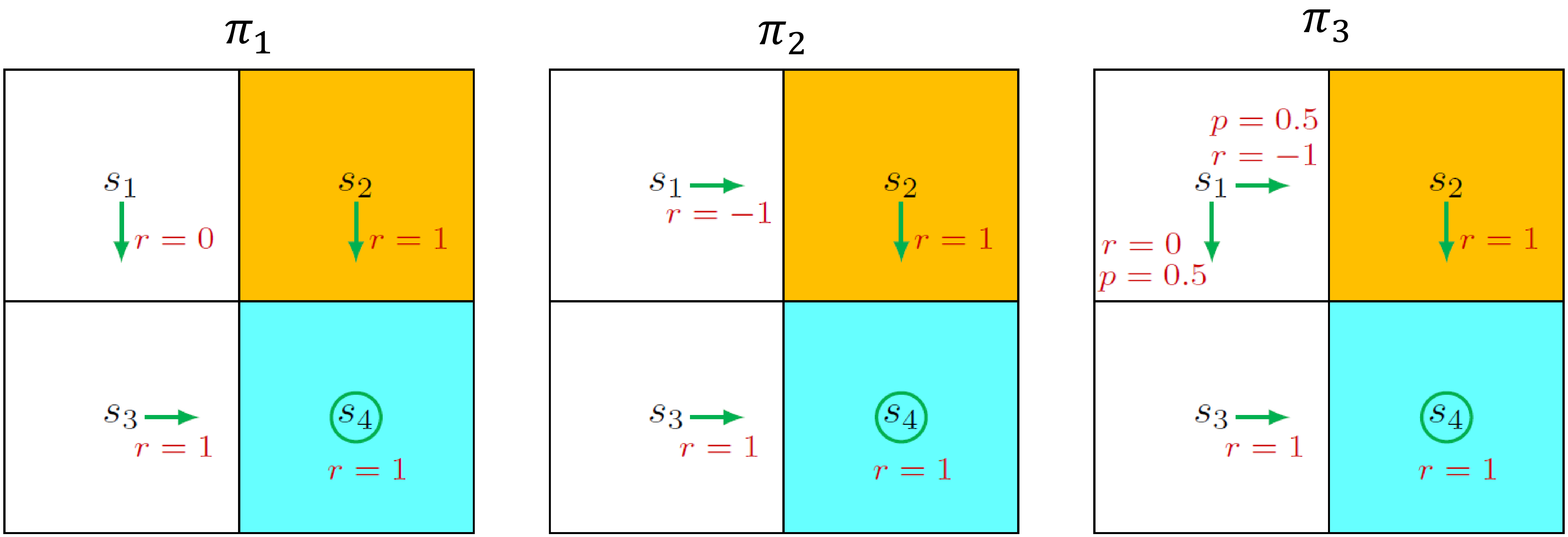
- Aprendizado por Reforço: interação agente-ambiente
  - O **agente** (tomador de decisão) percebe o estado, segue uma política e executa ações.
  - O **ambiente** inclui tudo aquilo que não é o agente e responde às suas ações.
  - Após executar uma ação, o estado do agente muda e ele recebe uma recompensa.

Valores de Estado e Equação de Bellman

- Conceito Fundamental: *Valor de estado*
  - *Recompensa média que um agente pode receber ao seguir uma determinada política.*
  - *Quanto maior o valor, melhor a política correspondente.*
  - *Usado para avaliar a qualidade de uma política.*
- Ferramenta Essencial: *Equação de Bellman*
  - Define relações entre os valores dos estados.
  - Usada para analisar os valores de estado.
  - Resolver a equação permite calcular os valores de estado (**avaliação de política**).

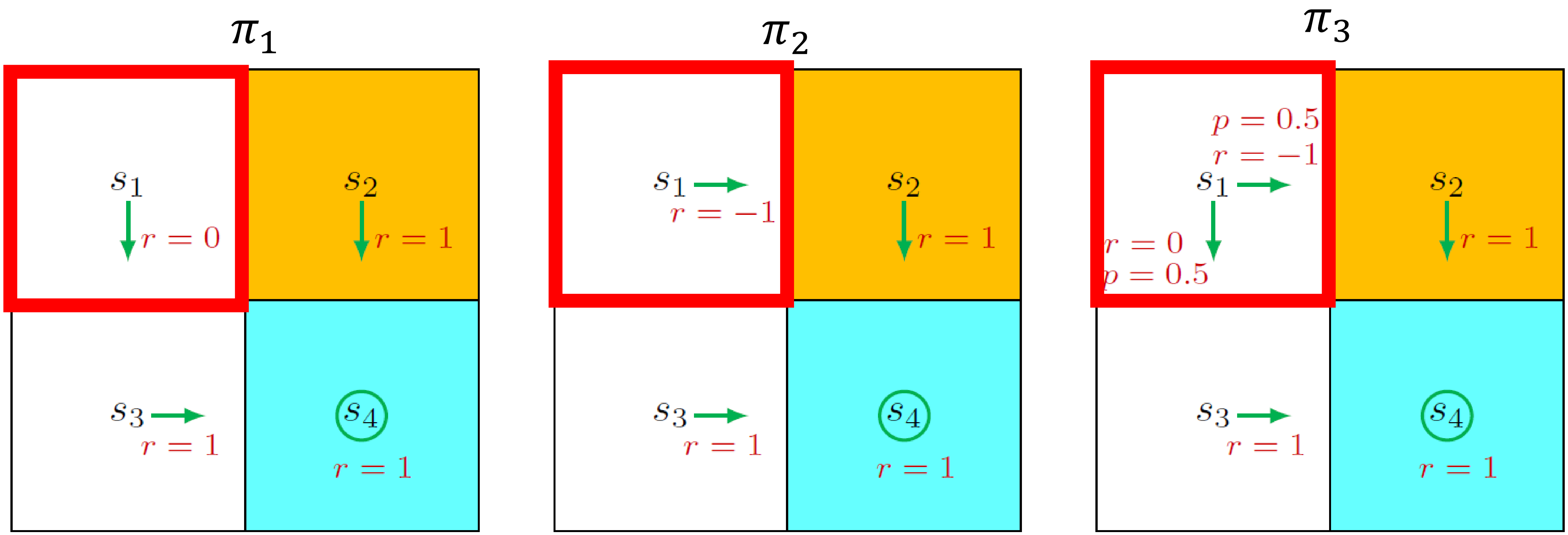
Valores de Estado e Equação de Bellman

- Qual a melhor política? π_1, π_2 ou π_3? E a pior? Por quê?



Valores de Estado e Equação de Bellman

- Qual a melhor política? π_1, π_2 ou π_3? E a pior? Por quê?



Valores de Estado e Equação de Bellman

- Assumindo estado inicial s_1 e taxa de desconto $\gamma \in (0,1)$

- Política 1:
  - Trajetória:

$s_1 \to s_3 \to s_4 \to s_4 \to \cdots$

  - Retorno descontado 1:

$$retorno_1 = 0 + \gamma 1 + \gamma^2 1 + \cdots$$

$$retorno_1 = \gamma(1 + \gamma + \gamma^2 + \cdots)$$

$$retorno_1 = \frac{\gamma}{1-\gamma}$$

- Política 2:
  - Trajetória:

$s_1 \to s_2 \to s_4 \to s_4 \to \cdots$

  - Retorno descontado 2:

$$retorno_2 = -1 + \gamma 1 + \gamma^2 1 + \cdots$$

$$retorno_2 = -1 + \gamma(1 + \gamma + \gamma^2 + \cdots)$$

$$retorno_2 = -1 + \frac{\gamma}{1-\gamma}$$

- Política 3:
  - Trajetórias

$s_1 \to s_3 \to s_4 \to s_4 \to \cdots$

$s_1 \to s_2 \to s_4 \to s_4 \to \cdots$

  - “Retorno” descontado 3:

$$retorno_3 = 0.5\left(-1 + \frac{\gamma}{1-\gamma}\right) + 0.5\left(\frac{\gamma}{1-\gamma}\right)$$

$$retorno_3 = -0.5 + \frac{\gamma}{1-\gamma}$$

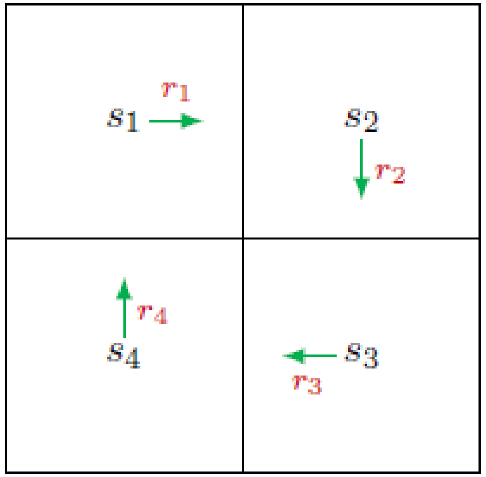
$$\boxed{retorno_1 > retorno_3 > retorno_2}$$

Valores de Estado e Equação de Bellman

- Uso dos retornos para avaliação de políticas
  - Políticas podem ser comparadas com base nos retornos obtidos.
  - Uma política é melhor se gera um retorno maior.
- Observação sobre o $retorno_3$
  - Se assemelha mais a um valor esperado do que a definição de retorno convencional.
  - $retorno_3$ é na verdade um valor de estado!

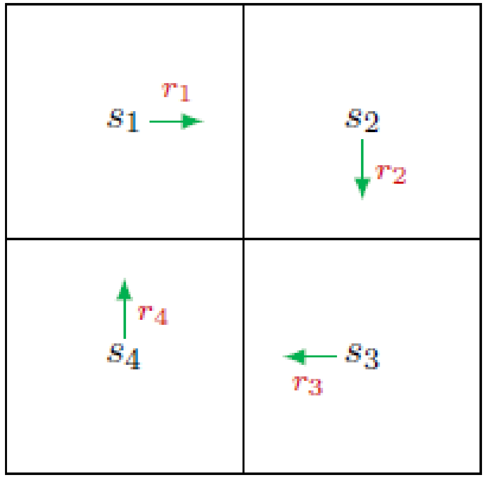
Valores de Estado e Equação de Bellman

- Cálculo dos retornos ao seguir uma política
  - Notação: v_i é o retorno obtido quando o estado inicial do agente é s_i $(i = 1,2,3,4)$.



Valores de Estado e Equação de Bellman

- Cálculo dos retornos ao seguir uma política
  - Usando a definição



$$v_1 = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots$$

$$v_2 = r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots$$

$$v_3 = r_3 + \gamma r_4 + \gamma^2 r_1 + \cdots$$

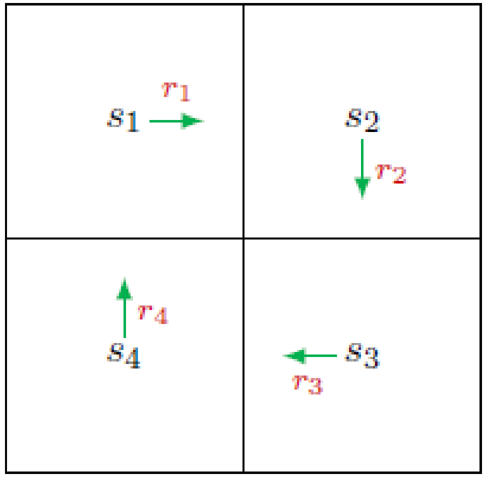
$$v_4 = r_4 + \gamma r_1 + \gamma^2 r_2 + \cdots$$

- Notação: v_i é o retorno obtido quando o estado inicial do agente é s_i $(i = 1,2,3,4)$.

Valores de Estado e Equação de Bellman

- Cálculo dos retornos ao seguir uma política
  - Usando o método de ***boostrapping*** (uso recursivo das próprias estimativas para calcular valores). Os retornos não são independentes; cada um deles depende do próximo.

Equação de Bellman



$$v_1 = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots$$

$$v_2 = r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots$$

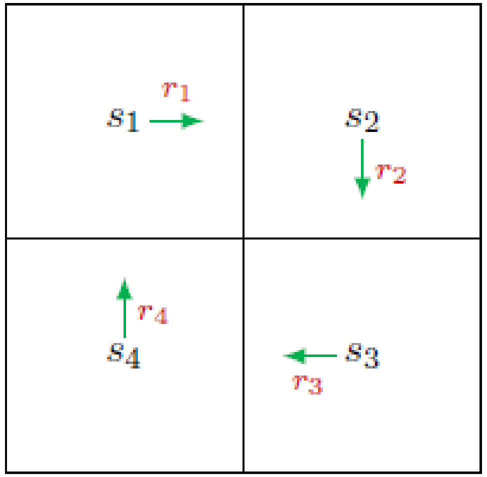
$$v_3 = r_3 + \gamma r_4 + \gamma^2 r_1 + \cdots$$

$$v_4 = r_4 + \gamma r_1 + \gamma^2 r_2 + \cdots$$

- Notação: v_i é o retorno obtido quando o estado inicial do agente é s_i $(i = 1,2,3,4)$.

Valores de Estado e Equação de Bellman

- Cálculo dos retornos ao seguir uma política
  - Usando o método de ***boostrapping*** (uso recursivo das próprias estimativas para calcular valores). Os retornos não são independentes; cada um deles depende do próximo.



$$v_1 = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots$$
$$v_2 = r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots$$
$$v_3 = r_3 + \gamma r_4 + \gamma^2 r_1 + \cdots$$
$$v_4 = r_4 + \gamma r_1 + \gamma^2 r_2 + \cdots$$

Equação de Bellman

$$v_1 = r_1 + \gamma(r_2 + \gamma r_3 + \cdots) = r_1 + \gamma v_2$$
$$v_2 = r_2 + \gamma(r_3 + \gamma r_4 + \cdots) = r_2 + \gamma v_3$$
$$v_3 = r_3 + \gamma(r_4 + \gamma r_1 + \cdots) = r_3 + \gamma v_4$$
$$v_4 = r_4 + \gamma(r_1 + \gamma r_2 + \cdots) = r_4 + \gamma v_1$$

  - Notação: v_i é o retorno obtido quando o estado inicial do agente é s_i $(i = 1,2,3,4)$.

Valores de Estado e Equação de Bellman

- Cálculo dos retornos ao seguir uma política
  - Usando o método de ***boostrapping***
    - Podemos usar notação matricial para o sistema de equações

$$v_1 = r_1 + \gamma(r_2 + \gamma r_3 + \cdots) = r_1 + \gamma v_2$$

$$v_2 = r_2 + \gamma(r_3 + \gamma r_4 + \cdots) = r_2 + \gamma v_3$$

$$v_3 = r_3 + \gamma(r_4 + \gamma r_1 + \cdots) = r_3 + \gamma v_4$$

$$v_4 = r_4 + \gamma(r_1 + \gamma r_2 + \cdots) = r_4 + \gamma v_1$$

$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{v} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{r} + \gamma \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{v}$$

    - De modo compacto: $v = r + \gamma P v$
    - Resolvendo para v: $v = (I - \gamma P)^{-1} r$

Valores de Estado e Equação de Bellman

- Limitações dos retornos na avaliação de políticas em sistemas estocásticos
  - O mesmo estado inicial pode levar a diferentes retornos.
  - Isso torna os retornos inadequados para avaliar políticas.
- Valor de Estado
  - O valor de estado é uma alternativa para lidar com a aleatoriedade nos retornos.
  - Representa a expectativa do retorno a partir de um estado específico.
- Notação
  - No tempo t, o agente está no estado S_t e executa a ação A_t seguindo a política π.
  - A transição leva ao próximo estado S_{t+1} e gera a recompensa imediata R_{t+1}.
  - Esse processo pode ser expresso como:

$$S_t \xrightarrow{A_t} S_{t+1}, R_{t+1}$$

Valores de Estado e Equação de Bellman

- 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

$$0 \xleftarrow{\text{"ênfase no curto prazo"}} \gamma \xrightarrow{\text{"ênfase no longo prazo"}} 1$$

Valores de Estado e Equação de Bellman

- Tarefas episódicas
  - Retorno sem taxa de desconto

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

$$G_t = \sum_{k=t+1}^{T} R_k$$

- Estado terminal ($\mathcal{S}^+$)

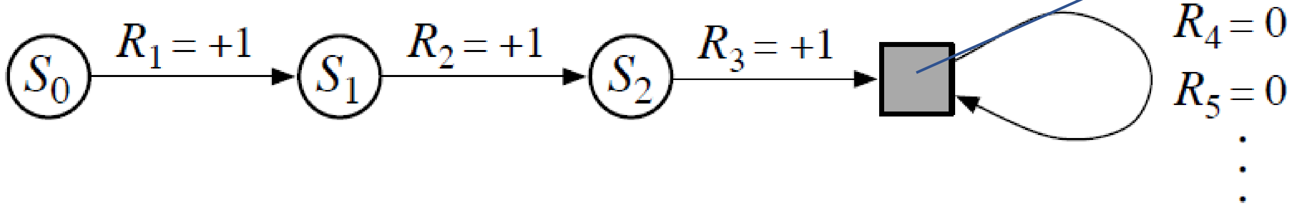
- Tarefas contínuas
  - Retorno com taxa de desconto

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

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{(t+1)+k}$$

- $T = \infty$

Estado absorvente



Valores de Estado e Equação de Bellman

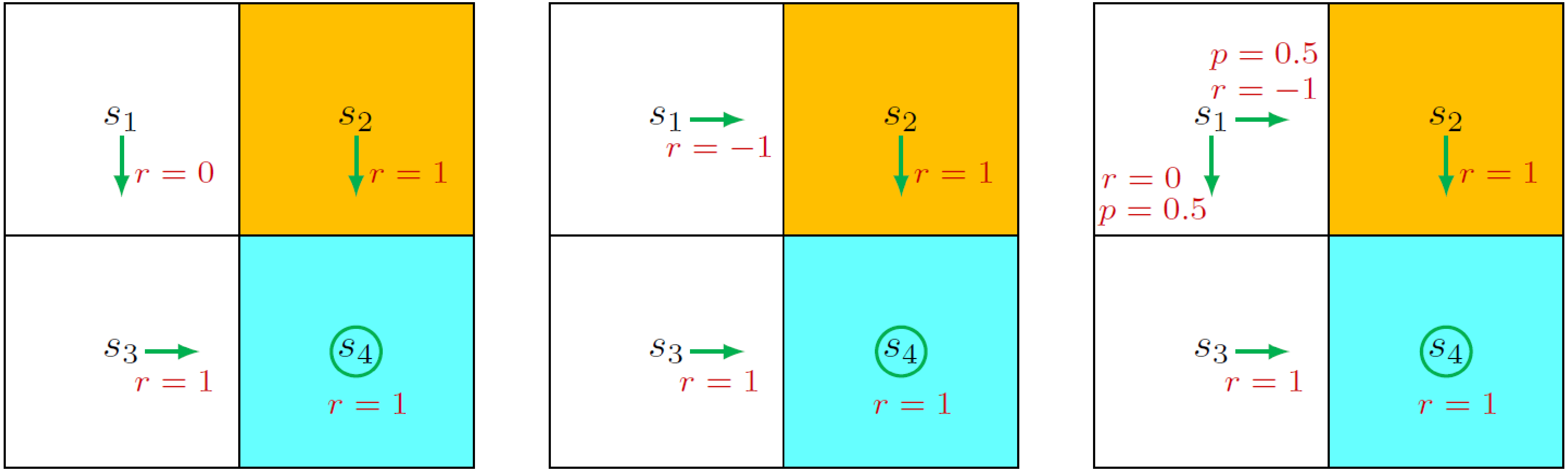
- Função de valor de estado (ou valor de estado de s) para a política π

$$v_\pi(s) \triangleq \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{(t+1)+k} \middle| S_t = s\right]$$

  - Depende do estado s, pois é um valor esperado condicional considerando que o estado inicial do agente é $S_t = s$.
  - Depende da política π, pois a trajetória do agente é gerada seguindo π.
- Relação entre Valores de Estado e Retornos
  - Se a política e o modelo do ambiente são **determinísticos:**
    - a trajetória partindo de um estado é sempre a mesma
    - Retorno = valor de estado.
  - Se a política ou o ambiente são **estocásticos:**
    - diferentes trajetórias podem ocorrer partindo do mesmo estado
    - O valor de estado passa a ser a média dos retornos possíveis.

Valores de Estado e Equação de Bellman

- Importância dos Valores de Estado na Avaliação de Políticas
  - O valor de estado é uma métrica mais formal para comparar políticas.
  - Políticas melhores são aquelas que geram valores de estado mais altos.



Referências

- Shiyu Zhao. Mathematical Foundations of Reinforcement Learning. Springer Singapore, 2025. [capítulos 1 e 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.