Aprendizado por Reforço

Métodos de Monte Carlo (parte 2)

Recapitulação da aula passada

- Aprendizado por reforço baseado em modelo
- Aprendizado por reforço sem modelo
- O problema da estimação das médias e sua relação com o aprendizado por reforço
- MC Básico

Recapitulação da aula passada

Algorithm 5.1: MC Basic (a model-free variant of policy iteration)

Initialization: Initial guess π_0.
Goal: Search for an optimal policy.

For the kth iteration ($k = 0, 1, 2, \ldots$), do
    For every state $s \in \mathcal{S}$, do
        For every action $a \in \mathcal{A}(s)$, do
            Collect sufficiently many episodes starting from (s, a) by following π_k
            Policy evaluation:
            $q_{\pi_k}(s, a) \approx q_k(s, a) =$ the average return of all the episodes starting from (s, a)
        Policy improvement:
        $a_k^*(s) = \arg\max_a q_k(s, a)$
        $\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise

Exemplo: Duração do Episódio e Recompensas Esparsas

- Mundo em grade de 5x5
- Recompensas definidas como:

$$r_{forbidden} = -10$$

$$r_{boundary} = -1$$

$$r_{target} = 1$$

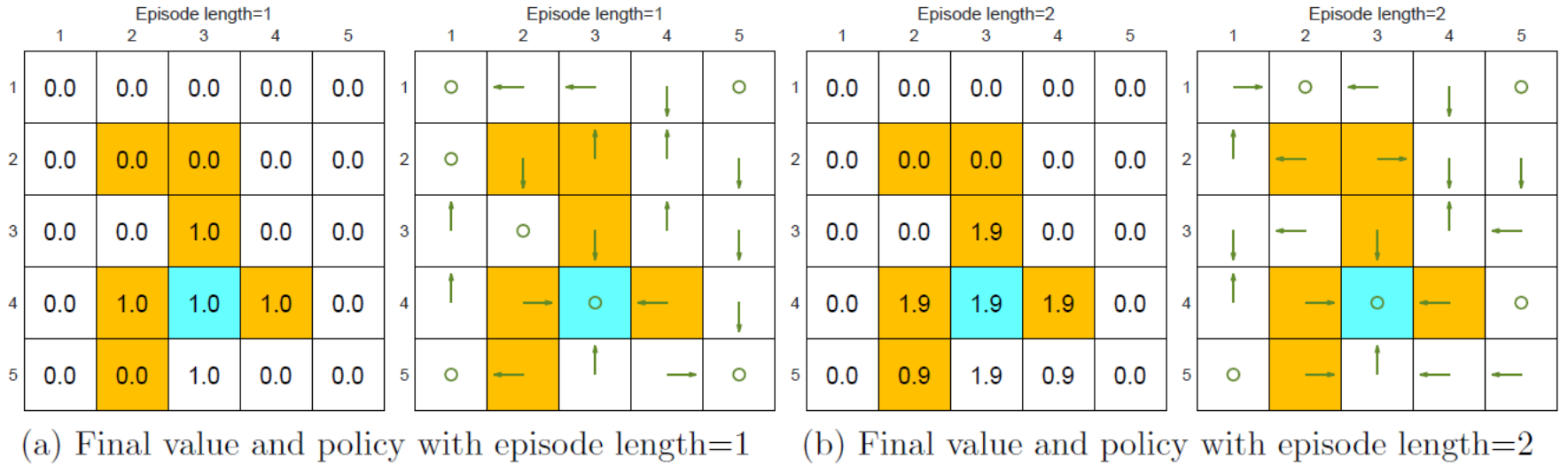
- Fator de desconto:

$$\gamma = 0.9$$

- Algoritmo: MC Básico

Exemplo: Duração do Episódio e Recompensas Esparsas

- Episódios curtos:
  - O que podemos observar?



(a) Final value and policy with episode length=1

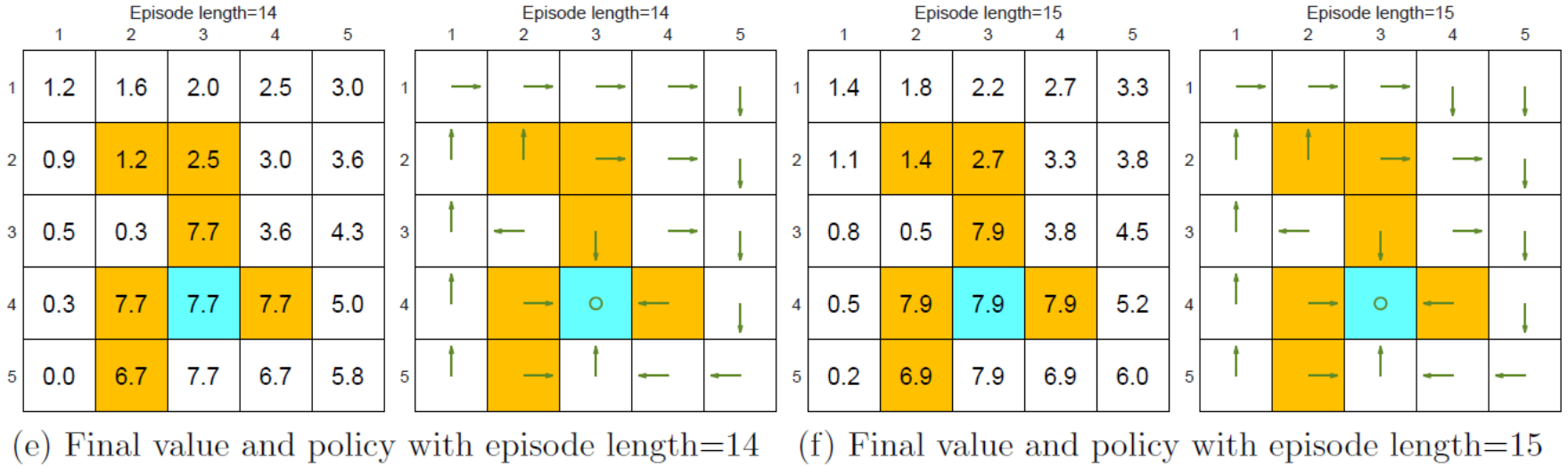
(b) Final value and policy with episode length=2

(c) Final value and policy with episode length=3

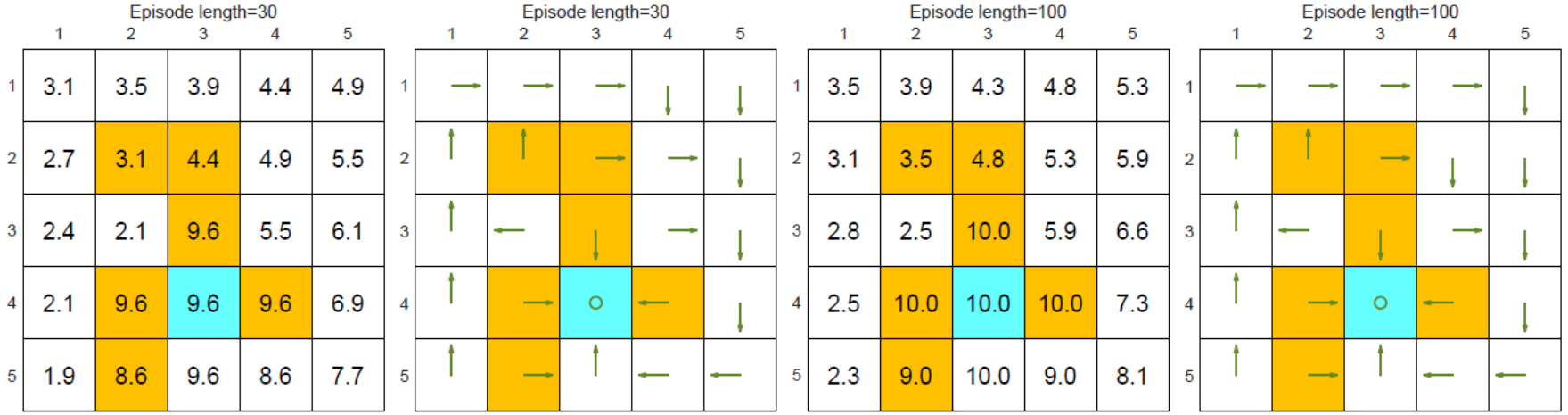
(d) Final value and policy with episode length=4

Exemplo: Duração do Episódio e Recompensas Esparsas

- Episódios longos:
  - O que podemos observar?



(e) Final value and policy with episode length=14 (f) Final value and policy with episode length=15



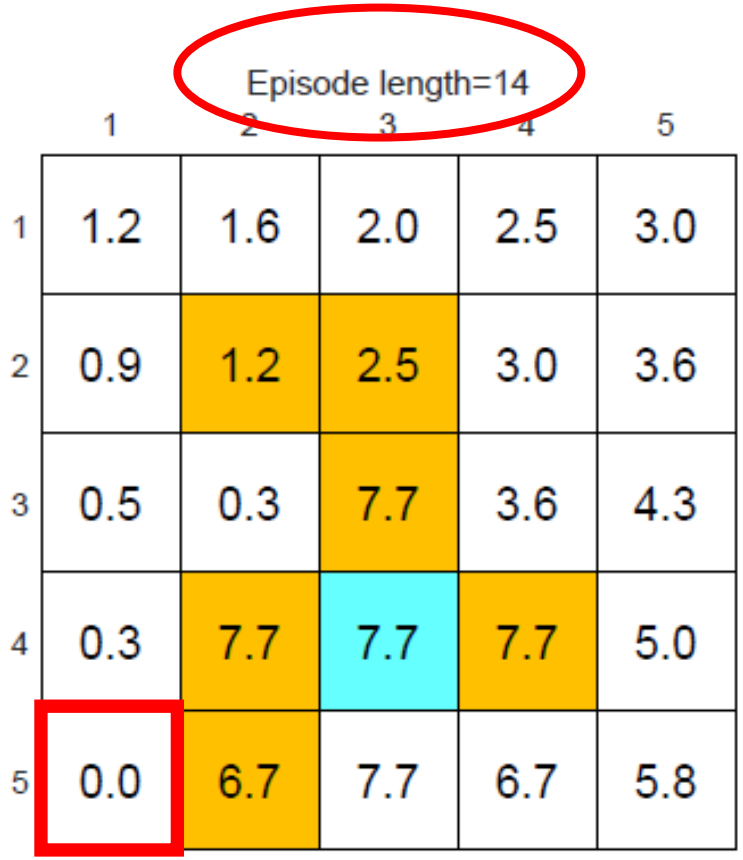
(g) Final value and policy with episode length=30 (h) Final value and policy with episode length=100

Efeito da Duração do Episódio

- A duração do episódio afeta fortemente a política ótima final.
- Episódios curtos ⟶ política e valores de estado não são ótimos.
  - Episódio com duração 1:
    - Apenas estados adjacentes ao alvo têm valor de estado diferente de zero.
    - Demais estados têm valor de estado zero.
- Episódios mais longos ⟶ aproximação gradual da política e valor de estado ótimos.

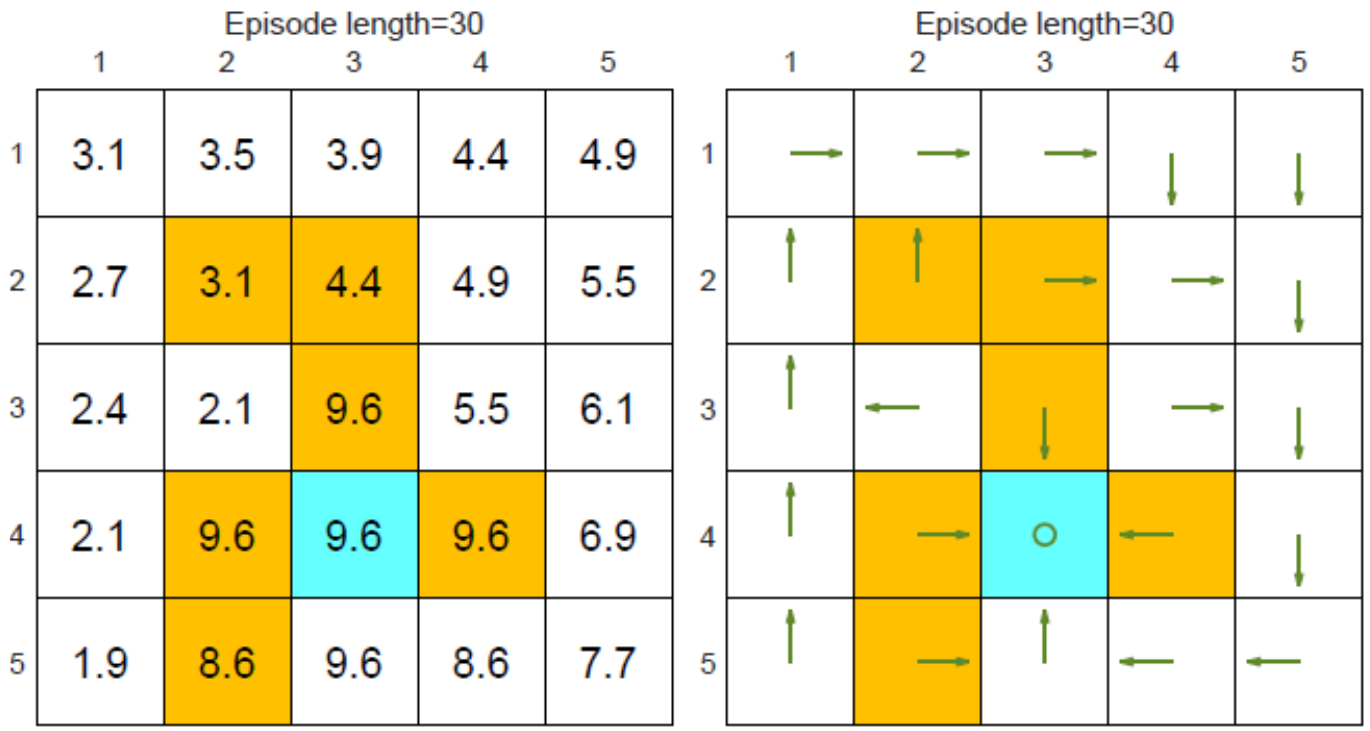
Padrões Espaciais Emergentes

- Com o aumento da duração dos episódios:
  - Estados próximos ao alvo ganham valor não-nulo antes dos estados mais afastados.
  - O agente precisa de um número mínimo de passos para alcançar o alvo.
    - Se episódio for mais curto que isso → retorno e valor do estado = 0.
  - Exemplo:
    - Do canto inferior esquerdo até o alvo → mínimo de 15 passos.
    - Logo, episódios devem ter pelo menos 15 passos.



Episódios Longos, Mas Não Infinitos

- Não é necessário que episódios sejam infinitos.
- Exemplo com episódio de 30 passos:
  - Valor de estado estimado ainda não é o ótimo.
  - Política ótima já é encontrada.



(g) Final value and policy with episode length=30

O Problema no projeto de recompensas

- Recompensas esparsas
  - Nenhuma recompensa positiva é obtida até atingir o alvo.
  - Exige episódios longos.
  - Dificulta o aprendizado em espaços de estado grandes.
  - Eficiência do aprendizado é reduzida.
- Recompensas não-esparsas
  - Estratégia: adicionar pequenas recompensas em estados próximos ao alvo.
  - Criação de um “campo atrativo” ao redor do alvo.
  - Agente encontra o alvo com maior facilidade.

Utilização de Amostras

- Suponha que tenhamos um episódio de amostras obtido ao seguir uma política π:

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \cdots$$

Como utilizar episódios de forma mais eficiente?

Utilização de Amostras

- **Visita:** ocorre toda vez que um par estado-ação (s, a) aparece em um episódio.
- Um episódio contém **várias visitas** a diferentes pares estado-ação.
- **MC Básico** adota a estratégia de **visita inicial** (*inicial-visit*):
  - Cada episódio é usado **apenas** para estimar o valor de ação **primeiro par** estado-ação (s, a) visitado.
  - Exemplo:
    - Episódio inicia em (s_1, a_2).
    - Episódio completo é usado apenas para estimar o valor de ação para (s_1, a_2):

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \cdots$$

- Estratégia com **baixo aproveitamento das amostras/ineficiente.**
  - O restante do episódio não é utilizado para fins de estimativa dos valores de ação para os outros pares.

Melhorando o aproveitamento das amostras

- Em um único episódio, diversos pares (s, a) são visitados.
- Podemos <u>decompor</u> o episódio em **subepisódios:**

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots$$ Episódio original

$$s_2 \xrightarrow{a_4} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots$$ Subepisódio iniciando em (s_2, a_4)

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots$$ Subepisódio iniciando em (s_1, a_2)

$$s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_1} \dots$$ Subepisódio iniciando em (s_2, a_3)

$$s_5 \xrightarrow{a_1} \dots$$ Subepisódio iniciando em (s_5, a_1)

- As visitas podem ser utilizadas para estimar os valores de ação correspondentes.
- Aumenta significativamente a eficiência no uso das amostras.

Melhorando o aproveitamento das amostras

- Um par estado-ação (s, a) pode ser visitado várias vezes em um episódio.

$$s_1 \overset{a_2}{\rightarrow} s_2 \overset{a_4}{\rightarrow} s_1 \overset{a_2}{\rightarrow} s_2 \overset{a_3}{\rightarrow} s_5 \overset{a_1}{\rightarrow} \cdots$$

- Primeira visita (*first-visit*):
  - Apenas a **primeira ocorrência** de cada par (s, a) é usada para estimação.
- Todas as visitas (*every-visit*) :
  - **Todas as ocorrências** de cada par (s, a) no episódio são utilizadas.
  - Um único episódio longo pode fornecer várias estimativas úteis.
  - Se o episódio for suficientemente longo e visitar vários pares muitas vezes, ele pode ser suficiente para estimar todos os valores de ação.
  - Estratégia **mais eficiente** em termos de uso de amostras.

Estratégias mais eficientes para atualizar a política

- **Atualizar política**: modificar as ações escolhidas com base nos novos valores de ação estimados.

1. Estratégia do **MC Básico** (durante a avaliação de política):
   - Agrega vários episódios iniciando no mesmo par (s, a).
   - Estima o valor de ação com a média dos retornos dos episódios.
   - Requer a espera da coleta de todos os episódios.
2. Estratégia do **MC Inícios Exploratórios**:
   - Usa o retorno de um único episódio para aproximar o valor da ação correspondente.
   - Permite obter uma estimativa imediata, assim que o episódio é recebido.
   - A política pode ser melhorada de forma incremental, episódio por episódio.
   - Mesmo sendo uma estimativa grosseira, já é suficiente para guiar atualizações.

Algoritmo MC com Inícios Exploratórios

- Combina o uso eficiente das amostras e atualizações frequentes de política.
- Utiliza a estratégia de **todas as visitas** (*every-visit*) para aproveitar melhor cada episódio.
- O cálculo do retorno descontado é feito de **trás para frente:**
  - Começa no estado final e retrocede até o par estado-ação inicial.
  - Essa abordagem aumenta a eficiência, mas também torna o algoritmo mais complexo.
- Condição de Inícios Exploratórios
  - Exige que haja múltiplos episódios iniciando em cada par estado-ação (s, a).
    - Isso garante estimativas confiáveis dos valores de ação, segundo a Lei dos Grandes Números.
  - Se um par não for suficientemente explorado:
    - Seu valor pode ser estimado de forma imprecisa.
    - A política pode não selecionar uma ação, mesmo que ela seja a melhor.

Algoritmo MC com Inícios Exploratórios

Algorithm 5.2: MC Exploring Starts (an efficient variant of MC Basic)

Initialization: Initial policy $\pi_0(a|s)$ and initial value $q(s,a)$ for all (s,a). Returns$(s,a) = 0$ and Num$(s,a) = 0$ for all (s,a).
Goal: Search for an optimal policy.

For each episode, do

Episode generation: Select a starting state-action pair (s_0, a_0) and ensure that all pairs can be possibly selected (this is the exploring-starts condition). Following the current policy, generate an episode of length T: $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$.
Initialization for each episode: $g \leftarrow 0$
For each step of the episode, $t = T-1, T-2, \ldots, 0$, do

$g \leftarrow \gamma g + r_{t+1}$
Returns$(s_t, a_t) \leftarrow$ Returns$(s_t, a_t) + g$
Num$(s_t, a_t) \leftarrow$ Num$(s_t, a_t) + 1$
Policy evaluation:
$q(s_t, a_t) \leftarrow$ Returns(s_t, a_t)/Num(s_t, a_t)
Policy improvement:
$\pi(a|s_t) = 1$ if $a = \arg\max_a q(s_t, a)$ and $\pi(a|s_t) = 0$ otherwise

Revisão

- **O que é estimação Monte Carlo?**
  - Estimação Monte Carlo refere-se a uma ampla classe de técnicas que utilizam amostras estocásticas para resolver problemas de aproximação.
- **O que é o problema de estimação da média?**
  - O problema de estimação da média refere-se ao cálculo do valor esperado de uma variável aleatória com base em amostras estocásticas.
- **Como resolver o problema de estimação da média?**
  - Abordagens:
  1. **Baseada em modelo:** se a distribuição de probabilidade de uma variável aleatória é conhecida, o valor esperado pode ser calculado com base em sua definição
  2. **Sem modelo:** se a distribuição de probabilidade é desconhecida, podemos usar estimação Monte Carlo para aproximar o valor esperado. Tal aproximação é precisa quando o número de amostras é grande.

Revisão

- **Por que o problema de estimação da média é importante para o aprendizado por reforço?**
  - Tanto os valores de estado quanto os valores de ação são definidos como valores esperados dos retornos. Portanto, estimar valores de estado ou de ação é essencialmente um problema de estimação da média.
- **Qual é a ideia central do aprendizado por reforço baseado em Monte Carlo sem modelo?**
  - A ideia central é converter o algoritmo de iteração de política em um algoritmo sem modelo.
    - Enquanto o algoritmo de iteração de política visa calcular valores com base no modelo do sistema, o aprendizado por reforço baseado em Monte Carlo substitui a etapa de avaliação de política baseada em modelo pela etapa de avaliação de política baseada em MC sem modelo.
- **O que são as estratégias de visita inicial, primeira visita e todas as visitas?**
  - São estratégias diferentes para utilizar as amostras de um episódio. Um episódio pode visitar muitos pares estado-ação.
    - **Visita inicial:** usa o episódio inteiro para estimar o valor de ação do par estado-ação inicial.
  - As estratégias de todas as visitas e de primeira visita podem aproveitar melhor as amostras disponíveis:
    - **Todas as visitas**: o restante do episódio é usado para estimar o valor de ação de um par estado-ação toda vez que ele é visitado.
    - **Primeira visita**: consideramos apenas a primeira vez que um par estado-ação é visitado no episódio.

Referências

- Shiyu Zhao. Mathematical Foundations of Reinforcement Learning. Springer Singapore, 2025. [capítulo 5]
  - 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 5]
  - 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.