Aprendizado por Reforço

Métodos de Monte Carlo (parte 1)

Recapitulação das aulas passadas...

- Algoritmos de aprendizado por reforço **baseados em modelos** (*model-based*)
  - Algoritmos de programação dinâmica para encontrar políticas ótimas
    - ✓Iteração de valor
    - ✓Iteração de política
    - ✓Iteração de política truncada

Mas, e se o modelo do ambiente não estiver disponível? Como encontrar políticas ótimas?

Métodos de Monte Carlo

Se não possuirmos acesso a um modelo do ambiente, **Então precisamos de dados**.

Se não possuirmos acesso a dados, **Então precisamos de um modelo do ambiente**.

Dados no contexto de aprendizado por reforço = interação do agente com o ambiente (experiência).

Métodos de Monte Carlo

- Classe de técnicas que usam amostras para resolver problemas de estimação.

O problema da estimação das médias

- Considere uma variável aleatória X com suporte $\mathcal{X}$ (um conjunto finito de números reais)
- Podemos calcular o valor esperado de X de duas maneiras:

1. Abordagem baseada em modelo:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x)\, x$$

Modelo

2. Abordagem sem modelo:

$$\mathbb{E}[X] \approx \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Amostras de X: $\{x_1, \ldots, x_n\}$

O problema da estimação das médias

- Na abordagem sem modelo, se o número de amostras n for pequeno, a aproximação pode não ser precisa.

- Entretanto, quando $n \to \infty$, temos que $\bar{x} \to \mathbb{E}[X]$ (**lei dos grandes números**).

O problema da estimação das médias

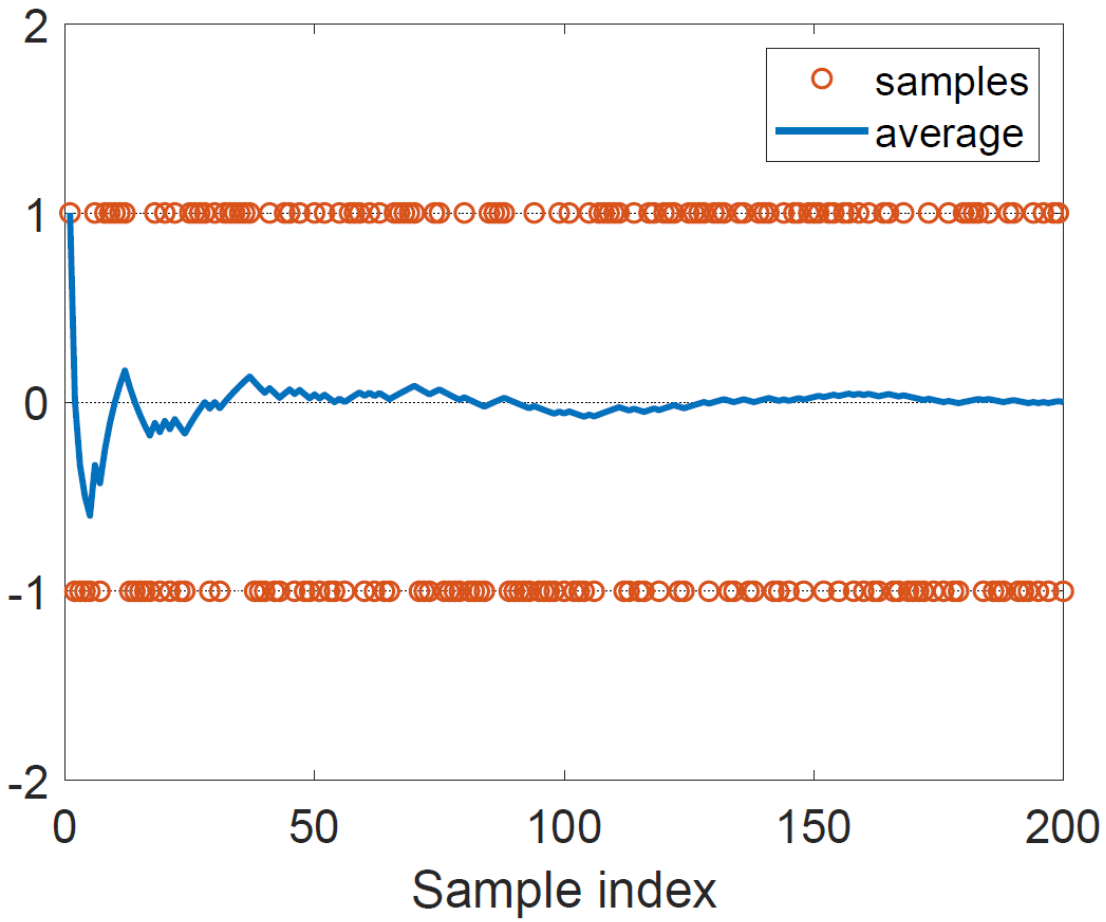
- Considere o lançamento de uma moeda honesta (não-viesada):
  - $\mathcal{X} = \{cara, coroa\}$
  - $p(X = cara) = p(X = coroa) = 0.5$
  - Seja $cara = 1$ e $coroa = -1$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x)x = 0.5(1) + 0.5(-1) = 0$$

O problema da estimação das médias

- Se $p(X = cara)$ e $p(X = coroa)$ forem desconhecidos, podemos lançar a moeda muitas vezes, amostrar os resultados $\{x_i\}_{i=1}^{n}$ e então estimar o valor esperado de X.

- Suposição:
  - Amostras são independentes e identicamente distribuídas (i.i.d)



Métodos de Monte Carlo

Por que nos importamos com problemas de estimação de médias?

Métodos de Monte Carlo

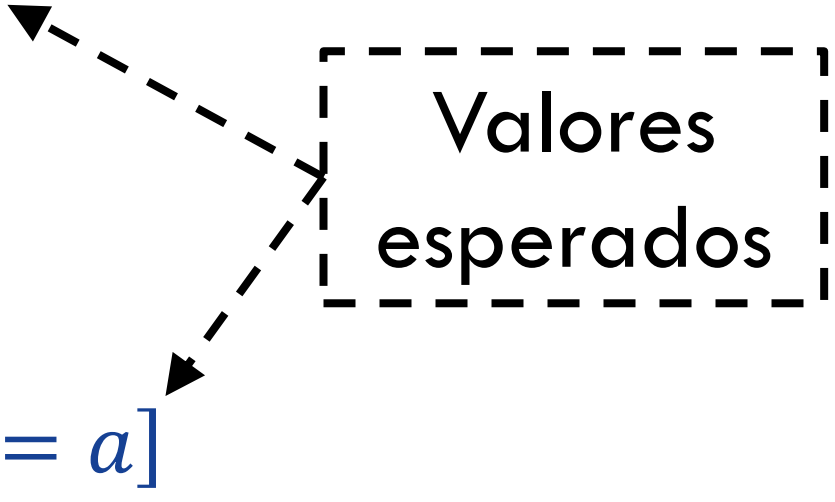
- Por que nos importamos com problemas de estimação de médias?

- Definição de valor de estado

$$v_\pi(s) \triangleq \mathbb{E}_\pi[G_t | S_t = s]$$

- Definição de valor de ação

$$q_\pi(s, a) \triangleq \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$



MC Básico

- Substituímos o passo de avaliação de política baseada em modelo no algoritmo de iteração de política por um **passo de estimação de MC sem modelo**.

Conversão da iteração de política para uma abordagem sem modelo

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

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

  - Passo 2: Melhoria de política
    - Forma matricial

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

    - Forma escalar

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

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

MC Básico

- Os **valores de ação** são os elementos centrais desses 2 passos.
  - Passo 1:
    - Calculamos os valores de estado para podermos calcular os **valores de ação**.
  - Passo 2:
    - Geramos a nova política utilizando os **valores de ação** calculados.

MC Básico

- Maneiras de calcular os valores de ação
  1. Abordagem baseada em modelo
     - Utilizada pelo algoritmo de iteração de política.
     - Após calcular os valores de estado (v_{π_k}), calculamos os valores de ação como:

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

- **Requisito**: o modelo do ambiente $\{p(r|s,a), p(s'|s,a)\}$ deve ser conhecido.

MC Básico

- Maneiras de calcular os valores de ação
  2. Abordagem sem modelo
     - Definição de valor de ação

$$q_{\pi_k}(s,a) \triangleq \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a]$$

Valor esperado do retorno quando se inicia em (s, a)

$$q_{\pi_k}(s,a) \triangleq \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \ | S_t = s, A_t = a]$$

- $q_{\pi_k}(s,a)$ é um valor esperado, então podemos estimá-lo utilizando **métodos de Monte Carlo**.

MC Básico

- Maneiras de calcular os valores de ação
  2. Abordagem sem modelo
     - Iniciando em (s, a) o agente pode interagir com o ambiente seguindo a política π_k de modo a obter um número n de episódios.
     - Seja $g_{\pi_k}^{(i)}(s, a)$ o retorno do i-ésimo episódio, podemos aproximar $q_{\pi_k}(s, a)$ como:

$$q_{\pi_k}(s, a) \triangleq \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a]$$

$$q_{\pi_k}(s, a) \approx \frac{1}{n} \sum_{j=1}^{n} g_{\pi_k}^{(i)}(s, a)$$

     - Pela lei dos grandes números, se n for suficientemente grande, então a aproximação será suficientemente precisa.

MC Básico

- Dada uma política inicial π_0, temos 2 passos na k-ésima iteração
  - Passo 1: Avaliação de política
    - Coletamos muitos episódios e usamos seus retornos para calcular $q_k(s,a)$, que é a aproximação de $q_{\pi_k}(s,a)$ via método de Monte Carlo.
  - Passo 2: Melhoria de política

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

- A política ótima (**gulosa**) é:

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

MC Básico

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

MC Básico vs. Iteração de política

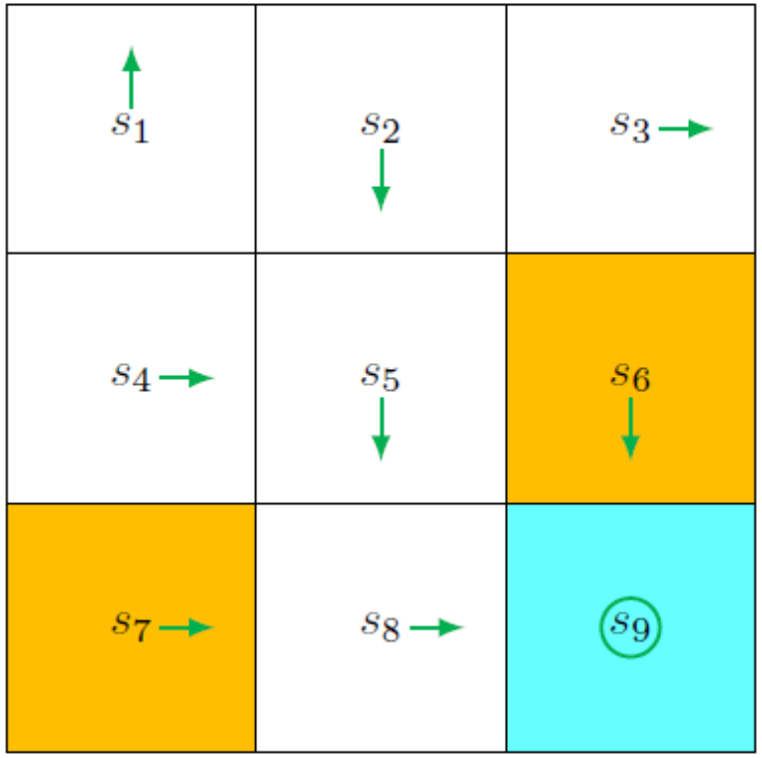
- MC Básico
  - Os valores de ação do são obtidos das amostras (experiência).
  - Estima diretamente os valores de ação (se estimássemos os valores de estado precisaríamos do modelo para estimar os valores de ação).
- Iteração de política
  - Calcula primeiro os valores de estado e depois os valores de ação baseado no modelo.

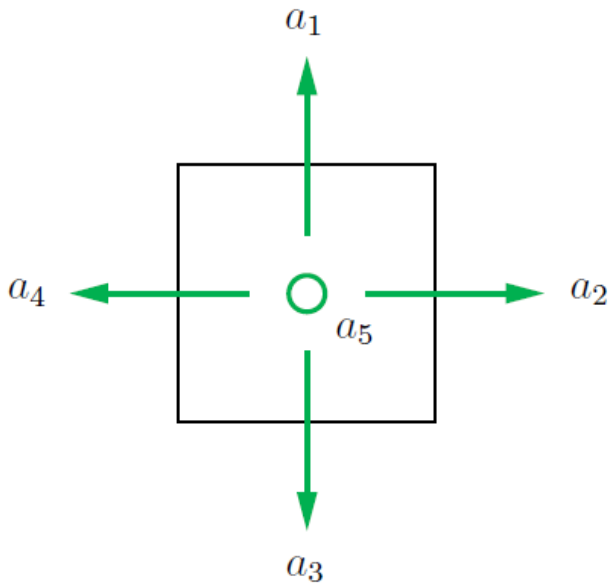
MC Básico

- Convergência
  - Assim como na iteração de política, o MC Básico também converge quando fornecido um número suficiente de amostras.
- Precisão na estimação de valores de ação
  - Para cada par (s, a), se houver episódios suficientes iniciando em (s, a), a média dos retornos desses episódios aproxima com precisão o valor de ação de (s, a).
- Na prática:
  - Frequentemente, **não** dispomos de episódios suficientes para cada (s, a).
  - Como consequência, as estimativas dos valores de ação podem ser imprecisas.
  - Mesmo assim, o algoritmo costuma funcionar adequadamente.
- Semelhança com a iteração de política truncada:
  - Em ambos os casos, os valores de ação são apenas aproximados, não exatos.
- Limitação:
  - O MC Básico é pouco eficiente no uso de amostras, o que não o torna prático.

MC Básico

- Exemplo





$r_{forbidden} = -1$

$r_{boundary} = -1$

$r_{target} = 1$

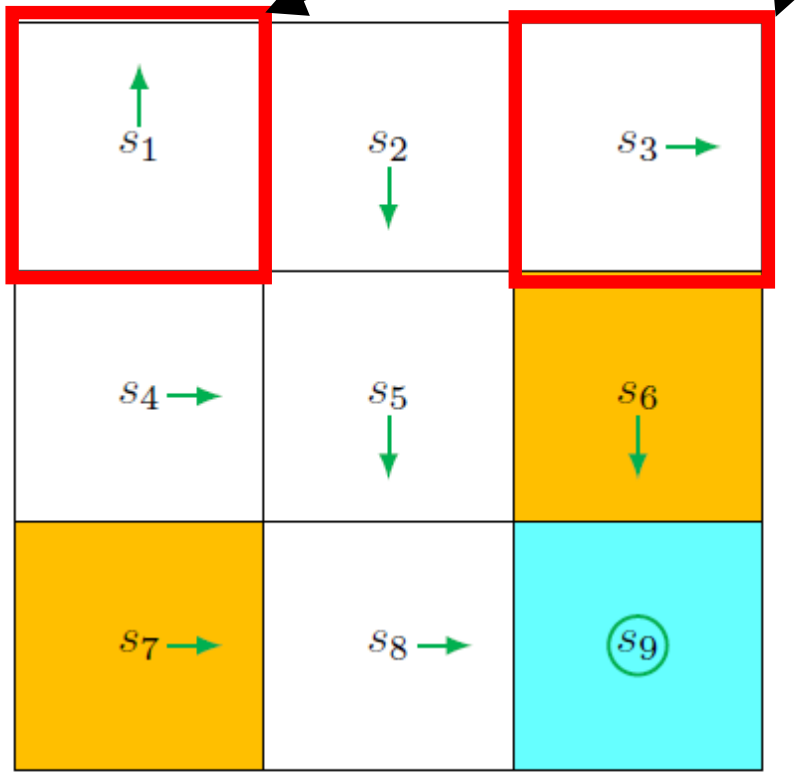
$\gamma = 0.9$

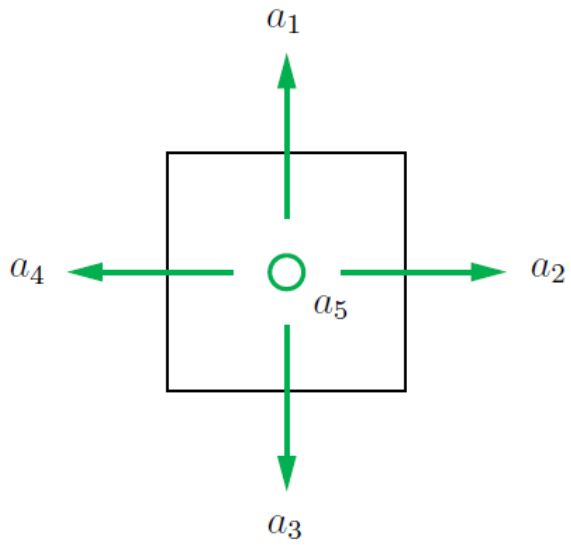
Política inicial: π_0

MC Básico

- Exemplo

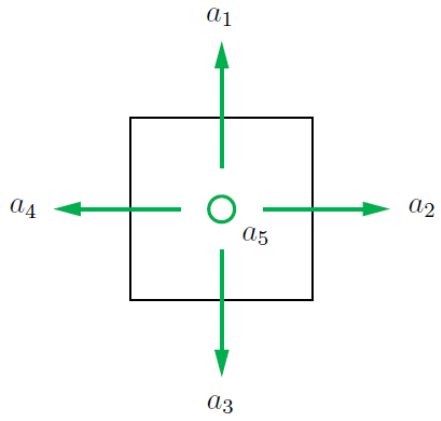
A política π_0 não é ótima para s_1 e s_3.

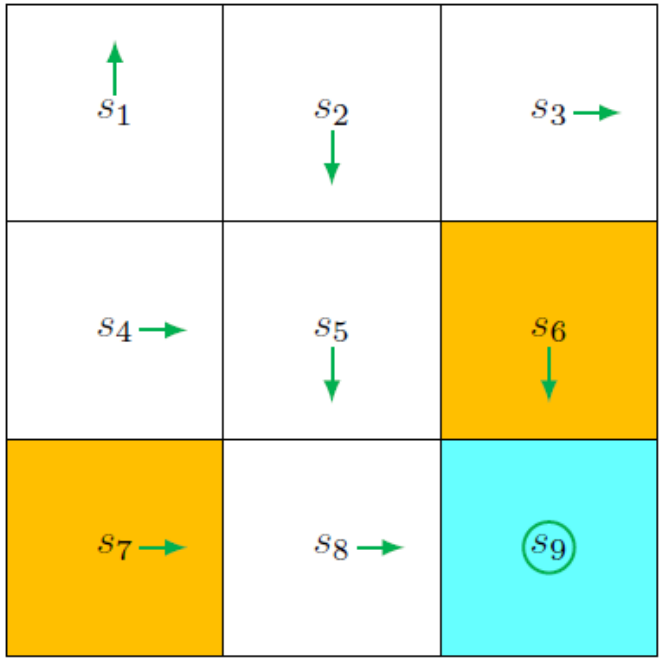




- Em s_1, existem 5 ações possíveis:
  - Para cada ação, precisamos coletar muitos episódios que sejam suficientemente longos para efetivamente **aproximar o valor de ação**.
- Este exemplo é **determinístico** (modelo & política), executar múltiplas vezes gera a mesma trajetória.
  - A estimação de cada valor de ação requer apenas **um único episódio**.

MC Básico - Exemplo





Seguindo π_0 e iniciando em

Início	Episódio	Valor de ação = retorno descontado do episódio
(s_1, a_1)	$s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \cdots$	$q_{\pi_0}(s_1, a_1) = -1 + \gamma(-1) + \gamma^2(-1) + \cdots = \frac{-1}{1-\gamma}$
(s_1, a_2)	$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_5 \xrightarrow{a_3} \cdots$	$q_{\pi_0}(s_1, a_2) = 0 + \gamma(0) + \gamma^2(0) + \gamma^3(1) + \gamma^4(1) \cdots = \frac{\gamma^3}{1-\gamma}$
(s_1, a_3)	$s_1 \xrightarrow{a_3} s_4 \xrightarrow{a_2} s_5 \xrightarrow{a_3} \cdots$	$q_{\pi_0}(s_1, a_3) = 0 + \gamma(0) + \gamma^2(0) + \gamma^3(1) + \gamma^4(1) \cdots = \frac{\gamma^3}{1-\gamma}$
(s_1, a_4)	$s_1 \xrightarrow{a_4} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \cdots$	$q_{\pi_0}(s_1, a_4) = -1 + \gamma(-1) + \gamma^2(-1) + \cdots = \frac{-1}{1-\gamma}$
(s_1, a_5)	$s_1 \xrightarrow{a_5} s_1 \xrightarrow{a_1} s_1 \xrightarrow{a_1} \cdots$	$q_{\pi_0}(s_1, a_5) = 0 + \gamma(-1) + \gamma^2(-1) + \cdots = \frac{-\gamma}{1-\gamma}$

MC Básico - Exemplo

- Comparando os 5 valores de ação $q_{\pi_0}(s_1, \cdot)$:

$$q_{\pi_0}(s_1, a_2) = q_{\pi_0}(s_1, a_3) = \frac{\gamma^3}{1-\gamma} > 0$$

- A nova política pode ser obtida como:

$$\pi_1(a_2|s_1) = 1 \quad \text{ou} \quad \pi_1(a_3|s_1) = 1$$

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.