Aprendizado por Reforço

Aproximação estocástica – parte 1

Introdução

- Métodos baseados em modelo vs. Métodos sem modelo.
- Métodos não-incrementais vs. Métodos incrementais.
  - Exemplo:
    - Monte Carlo (**não-incremental**)
    - Aprendizado por diferença temporal (**incremental**)
      - **Algoritmos de diferença temporal** podem ser vistos como casos especiais de **aproximação estocástica**.

Estimação da média

- Problema:
  - Estimar $\mathbb{E}[X]$, onde X é uma variável aleatória cujo suporte S_X é um conjunto finito.
  - Temos uma sequência de amostras i.i.d.: $\{x_i\}_{i=1}^n$.
  - Aproximação:

$$\mathbb{E}[X] \approx \bar{x} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estimativa de Monte Carlo

  - Pela lei dos grandes números: $\bar{x} \to \mathbb{E}[X]$ quando $n \to \infty$.

Estimação da média

- Métodos para calcular $\bar{x}$:
  - Não-incremental: calcula diretamente a média das $k+1$ amostras.

$$w_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i$$

  - **Como converter o método não-incremental em um incremental?**

Estimação da média

- **Como converter o método não-incremental em um incremental?**

$$w_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i\,, \qquad k = 1, 2, \ldots$$

$$w_k = \frac{1}{k}\sum_{i=1}^{k} x_i\,, \qquad k = 1, 2, \ldots$$

Estimação da média

- **Como converter o método não-incremental em um incremental?**

- Reescrevemos:

$$w_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i = \frac{1}{k+1}\left(\sum_{i=1}^{k} x_i + x_{k+1}\right) = \frac{1}{k+1}(kw_k + x_{k+1})$$

$$w_{k+1} = \frac{1}{k+1}\left((k+1-1)w_k + x_{k+1}\right) = \frac{1}{k+1}\left((k+1)w_k - w_k + x_{k+1}\right)$$

$$\boxed{w_{k+1} = w_k - \frac{1}{k+1}(w_k - x_{k+1})}$$

Estimação da média

- Verificação:

$$w_1 = x_1$$

$$w_2 = w_1 - \frac{1}{2}(w_1 - x_2) = \frac{1}{2}(x_1 + x_2)$$

$$w_3 = w_2 - \frac{1}{3}(w_2 - x_3) = \frac{1}{3}(x_1 + x_2 + x_3)$$

$$\vdots$$

$$w_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i$$

$$w_2 = w_1 - \frac{1}{2}(w_1 - x_2) = x_1 - \frac{1}{2}(x_1 - x_2)$$

$$w_2 = \left(1 - \frac{1}{2}\right)x_1 + \frac{1}{2}x_2 = \boxed{\frac{1}{2}(x_1 + x_2)}$$

$$w_3 = w_2 - \frac{1}{3}(w_2 - x_3)$$

$$w_3 = \frac{1}{2}(x_1 + x_2) - \frac{1}{3}\left(\frac{1}{2}(x_1 + x_2) - x_3\right)$$

$$w_3 = \left(1 - \frac{1}{3}\right)\frac{1}{2}(x_1 + x_2) + \frac{1}{3}x_3$$

$$w_3 = \left(\frac{2}{3}\right)\frac{1}{2}(x_1 + x_2) + \frac{1}{3}x_3 = \boxed{\frac{1}{3}(x_1 + x_2 + x_3)}$$

Estimação da média

- Métodos para calcular $\bar{x}$:
  - **Não-incremental:**

$$w_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i$$

    - Desvantagem: necessário esperar a coleta de todas as amostras.
  - **Incremental**

$$w_{k+1} = w_k - \frac{1}{k+1}(w_k - x_{k+1})$$

    - **Vantagem:** permite atualização imediata da média com cada nova amostra x_{k+1}.
    - Inicialmente, a aproximação pode não ser precisa.
    - Com mais amostras, a precisão melhora pela lei dos grandes números.

Estimação da média

- Formulação geral:

$$w_{k+1} = w_k - \alpha_k(w_k - x_k)$$

- Igual à formulação incremental anterior, mas utilizando $\alpha_k > 0$.
- Sem a expressão de α_k, não podemos obter uma fórmula explícita para w_{k+1}.
- Se α_k satisfizer certas condições, então $w_{k+1} \to \mathbb{E}[X]$ quando $k \to \infty$.
- **Algoritmos de diferença temporal** possuem expressões semelhantes, porém mais complexas.

Algoritmo de Robbins-Monro

- **Aproximação estocástica:**
  - Classe de algoritmos iterativos estocásticos para resolver problemas de encontrar raízes ou de otimização.
  - Não requer a expressão da função objetivo nem de sua derivada.
  - O **algoritmo de Robbins-Monro** (RM) é um trabalho pioneiro área de aproximação estocástica.
  - O **algoritmo de descida do gradiente estocástico** é uma caso particular do algoritmo de Robbins-Monro.

Algoritmo de Robbins-Monro

- **Definição do problema:** encontrar a raiz da equação

$$g(w) = 0, \qquad w \in \mathbb{R}, \qquad g: \mathbb{R} \to \mathbb{R}$$

  - Problemas de otimização podem ser convertidos em problemas de localização de raiz. Seja $J(w)$ a função objetivo a ser otimizada, então

$$g(w) \triangleq \nabla_w J(w) = 0$$

  - Equações como $f(w) = c$ também podem ser convertidas ao se escrever

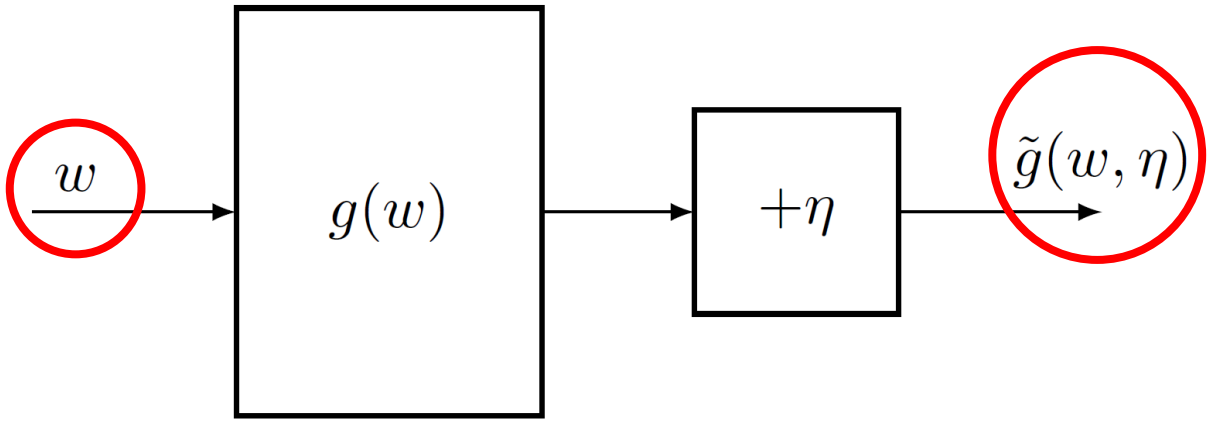
$$g(w) = f(w) - c$$

Algoritmo de Robbins-Monro

- Sistema de caixa-preta
  - Desconhecidos:
    - $g(w)$ e $\nabla_w g(w)$.
      - Se a forma analítica de g ou sua derivada fosse conhecida, poderíamos usar métodos numéricos.
  - Conhecidos
    - Entrada: w
    - Saída: $\tilde{g}(w, \eta)$

$$\tilde{g}(w, \eta) = g(w) + \eta, \qquad \eta \in \mathbb{R}$$

      - η: erro de observação
      - $\tilde{g}(w, \eta)$: observação ruidosa de $g(w)$.



- Objetivo:
  - Resolver $g(w) = 0$ usando w e $\tilde{g}(w, \eta)$.

Algoritmo de Robbins-Monro

- Algoritmo de Robbins-Monro que resolve $g(w) = 0$:

$$w_{k+1} = w_k - \alpha_k \tilde{g}(w_k, \eta_k), \qquad k = 1, 2, \ldots$$

  - w_k: k-ésima estimativa da raiz
  - $\tilde{g}(w_k, \eta_k)$: k-ésima observação ruidosa
  - α_k: coeficiente positivo
  - Não requer nenhuma informação sobre g, apenas entrada e saída.

Algoritmo de Robbins-Monro

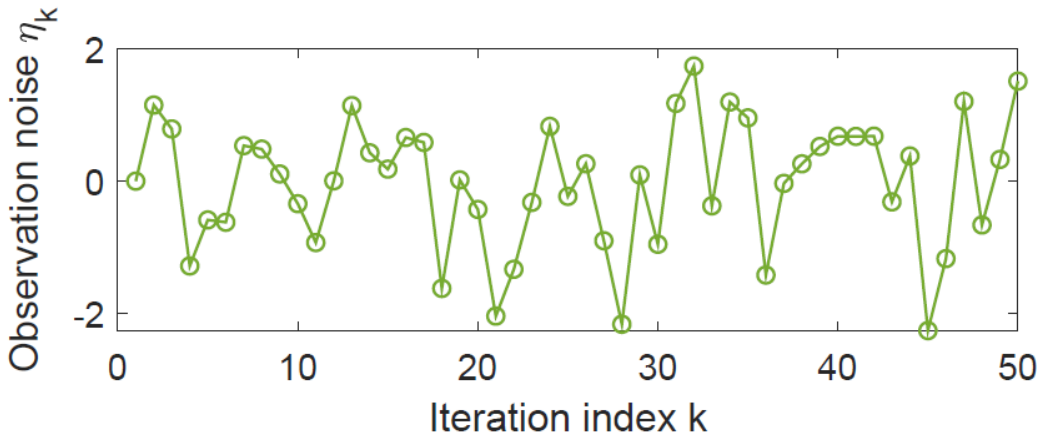
- Exemplo:

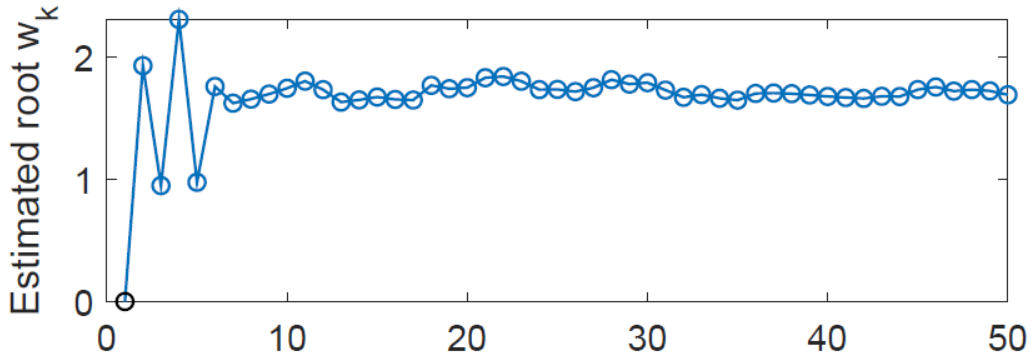
$$g(w) = w^3 - 5$$

- Raiz verdadeira: $5^{1/3} \approx 1.71$.
- Podemos observar somente:

$$\tilde{g}(w, \eta) = g(w) + \eta$$

  - Onde η é i.i.d. e $\eta \sim \mathcal{N}(0,1)$.
  - Inicialização: $w_1 = 0$.
  - Coeficiente: $a_k = 1/k$.
- Apesar do ruído, a as estimativas de w_k convergem para $5^{1/3} \approx 1.71$.
- Inicialização deve ser adequado para garantir convergência nesse exemplo.



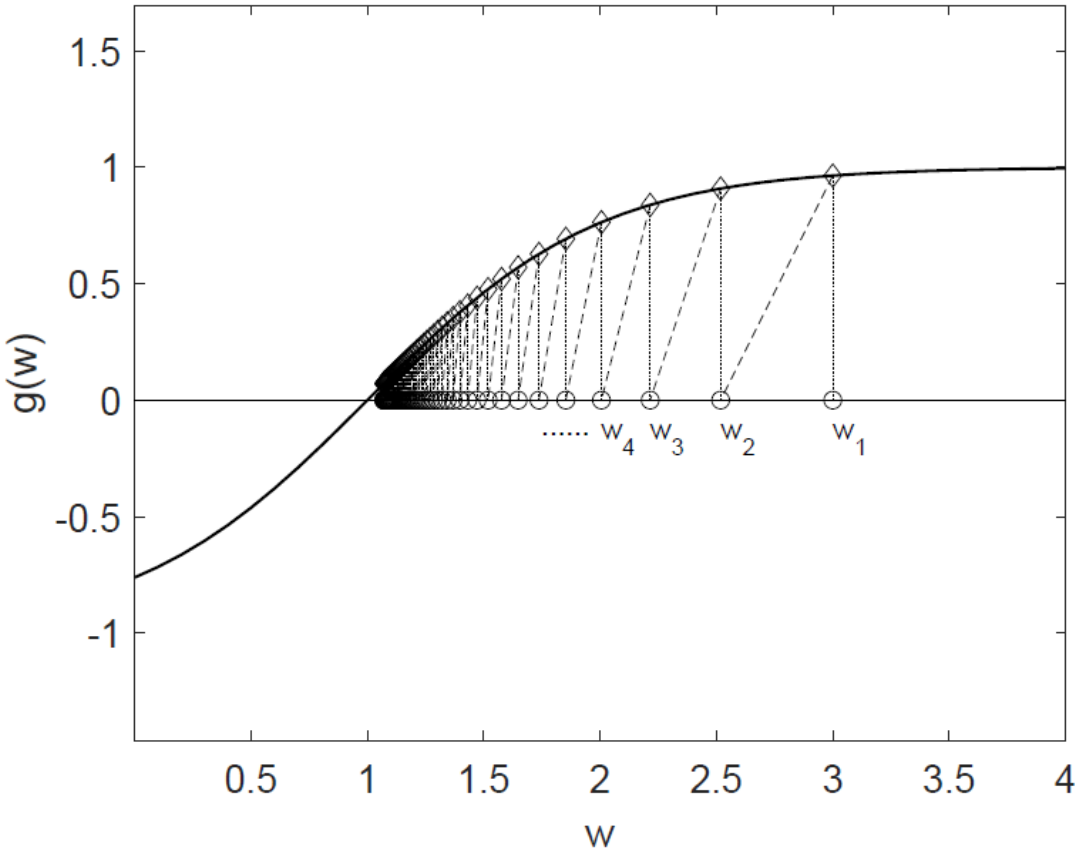


Algoritmo de Robbins-Monro

- Por que o algoritmo de Robbins-Monro encontra a raiz de $g(w) = 0$?
- Exemplo:
  - $g(w) = \tanh(w - 1)$ com raiz $w^* = 1$.
  - Algoritmo de Robbins-Monro:

$$w_{k+1} = w_k - \alpha_k g(w)$$

  - Inicialização: $w_1 = 3$
  - Coeficiente: $a_k = 1/k$
  - Considerando: $\eta \triangleq 0$



Algoritmo de Robbins-Monro

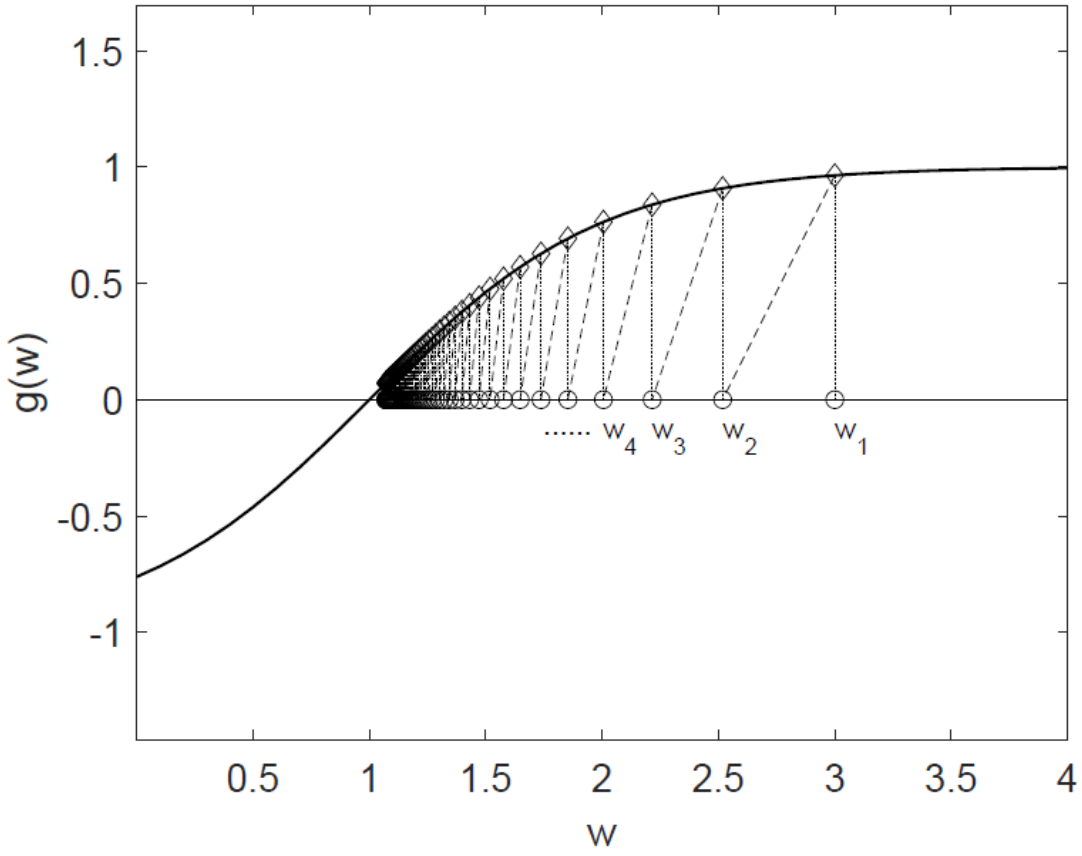
- Algoritmo de Robbins-Monro

$$w_{k+1} = w_k - \alpha_k g(w)$$

Considere os seguintes casos:

1. $w_k > w^* \rightarrow g(w_k) > 0 \rightarrow w_{k+1} = w_k - \alpha_k g(w_k) < w_k$
   - Se $\alpha_k g(w_k)$ for suficientemente pequeno: $w^* < w_{k+1} < w_k$
2. $w_k < w^* \rightarrow g(w_k) < 0 \rightarrow w_{k+1} = w_k - \alpha_k g(w_k) > w_k$
   - Se $|\alpha_k g(w)|$ for suficientemente pequeno: $w^* > w_{k+1} > w_k$

- Nos dois casos, w_{k+1} se aproxima de w^*.



Algoritmo de Robbins-Monro

- Teorema de Robbins-Monro

No Algoritmo de Robbins-Monro, se:

$a.\quad 0 < c_1 \leq \nabla_w g(w) \leq c_2 \ \forall w$

$b.\quad \sum_{k=1}^{\infty} a_k = \infty$ e $\sum_{k=1}^{\infty} a_k^2 < \infty$

$c.\quad \mathbb{E}[\eta_k | \mathcal{H}_k] = 0 \ e \ \mathbb{E}\left[a_k^1 | \mathcal{H}_k\right] < \infty$

Onde $\mathcal{H}_k = \{w_k, w_{k-1}, \ldots\}$, então $w_k \rightarrow w^*$ quase seguramente, com $g(w^*) = 0$.

Algoritmo de Robbins-Monro

- Aplicação do teorema de Robbins-Monro ao problema de estimação da média.
  - Algoritmo de estimação da média:

$$w_{k+1} = w_k - \alpha_k (w_k - x_k)$$

  - Para $\alpha_k = 1/k$, temos uma solução analítica.
  - Mas e para valores gerais de α_k?

Algoritmo de Robbins-Monro

Como reformular o problema de estimação da média como um problema de localização de raiz?

Algoritmo de Robbins-Monro

- Podemos reformular como um algoritmo de Robbins-Monro:

$$g(w) \triangleq w - \mathbb{E}[X]$$

  - Problema original: encontrar o valor de $\mathbb{E}[X] = w$.
  - Problema reformulado: resolver $g(w) = 0$ (encontrar a raiz).

Algoritmo de Robbins-Monro

- Observação ruidosa:

$$\tilde{g}(w,\eta) \triangleq w - x$$

- Note que:

$$\tilde{g}(w,\eta) = w - x = w - x + \mathbb{E}[X] - \mathbb{E}[X]$$

$$\tilde{g}(w,\eta) = (w - \mathbb{E}[X]) + (\mathbb{E}[X] - x)$$

$$\tilde{g}(w,\eta) = g(w) + \eta$$

Algoritmo de Robbins-Monro

- O algoritmo de Robbins-Monro para este problema é:

$$w_{k+1} = w_k - \alpha_k \tilde{g}(w_k, \eta_k)$$

$$\boxed{w_{k+1} = w_k - \alpha_k (w - x)}$$

- que é exatamente o algoritmo de estimação da média.
- Pelo Teorema de Robbins-Monro, $w_k \to \mathbb{E}[X]$ quase seguramente se $\sum_{k=1}^{\infty} a_k = \infty$, $\sum_{k=1}^{\infty} a_k^2 < \infty$ e $\{x_k\}$ for i.i.d. (independende da distribuição de X).

Referências

- S. Zhao. *Mathematical Foundations of Reinforcement Learning*. Springer Singapore, 2025. [capítulo 6]
  - **disponível em:** https://github.com/MathFoundationRL/Book-Mathematical-Foundation-of-Reinforcement-Learning

Slides construídos com base nos livros supracitados, os quais estão disponibilizados publicamente pelos seus respectivos autores.