INSTITUTO DE MATEMATICA E ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO

XLIX Programa de Verão (2020) - Introdução ao Aprendizado por Reforço

Policy Gradients

Ângelo Gregório Lovatto
aglovatto@ime.usp.br

IME - USP, 12/02/2019
LIAMF: Grupo PAR (Planejamento e Aprendizado por Reforço)

Aula 2

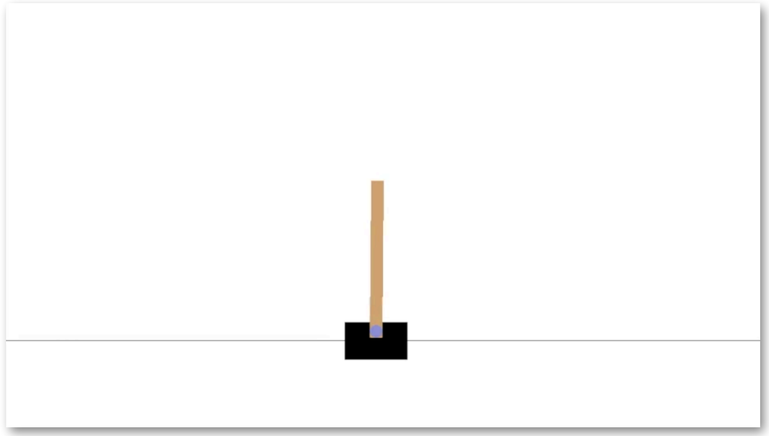
Agenda

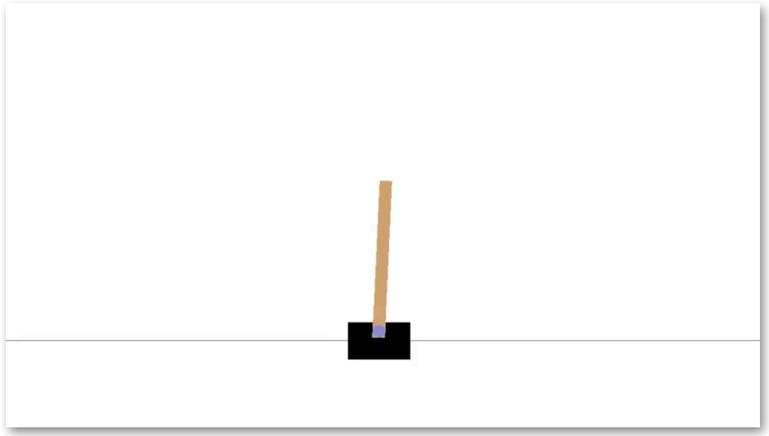
1. Introdução
2. Derivação do gradiente de política
3. Algoritmo (REINFORCE)

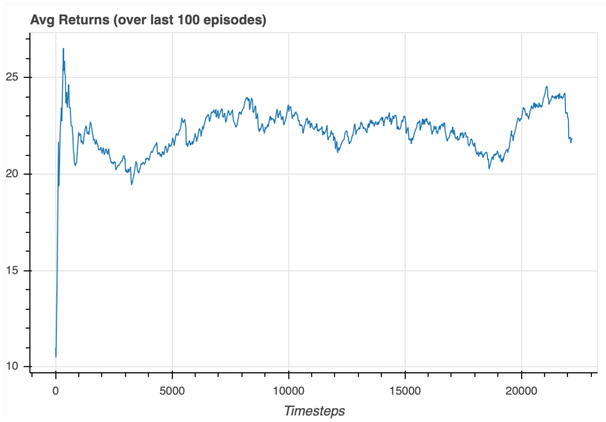
Objetivos

- Entender (model-free) RL como abordagem trial-and-error
- Entender abordagem de Policy Gradients como busca no espaço de parâmetros da política
- Implementar algoritmo REINFORCE via diferenciação automática

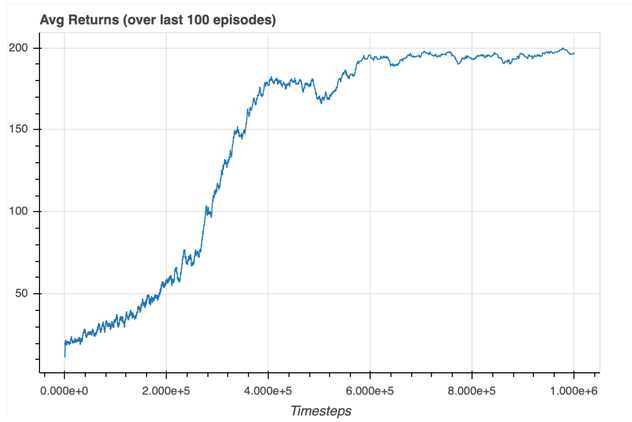
Aula de hoje





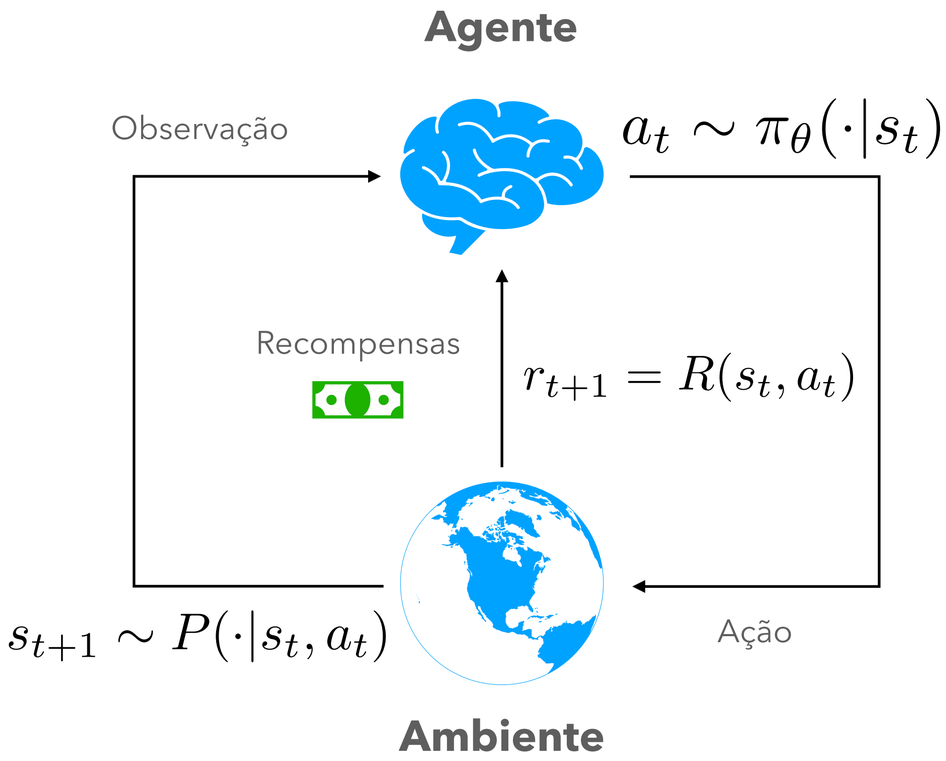


Policy Gradient



Motivação

- Como tomar ações melhores com base em experiências?
- Lembrar do ciclo de interação agente-ambiente
- As ações são escolhidas por uma política probabilística



Recapitulando

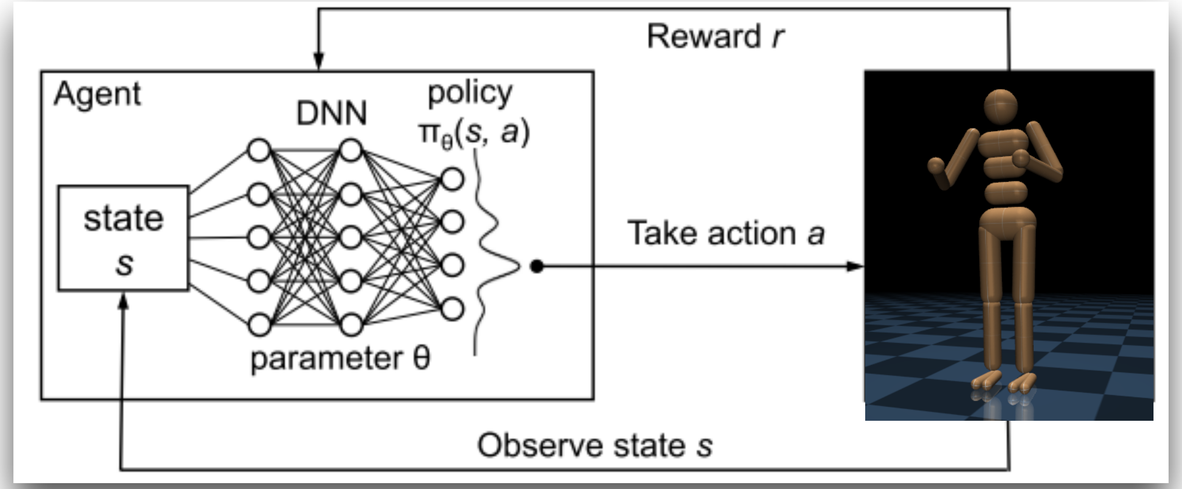
$$\tau = (s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_{T-1}, a_{T-1}, r_T, s_T)$$

$$R(\tau) = \sum_{t=0}^{T-1} r_{t+1}$$

$$\max_{\theta} J(\theta) = \max_{\theta} \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} [R(\tau)]$$

- Interações repetidas com o ambiente geram trajetórias
- O retorno de uma trajetória é a soma de suas recompensas
- Queremos maximizar o retorno esperado sob a política
- Como achar bons parâmetros para a política?

Policy Gradients - Como treinar um agente?



Mao et al. (2016). Resource Management with Deep Reinforcement Learning. HotNets, 50–56. ACM.

- RL envolve processar entradas sensoriais "cruas" e a partir disso recomendar uma ação
- Redes neurais tem se mostrado úteis para representar políticas complexas
- Métodos de gradiente descendente são altamente escaláveis
- Como aproveitar o maquinário de Deep Learning para otimizar tais políticas?

Otimização de política por gradientes

Suponha que temos uma política parametrizada

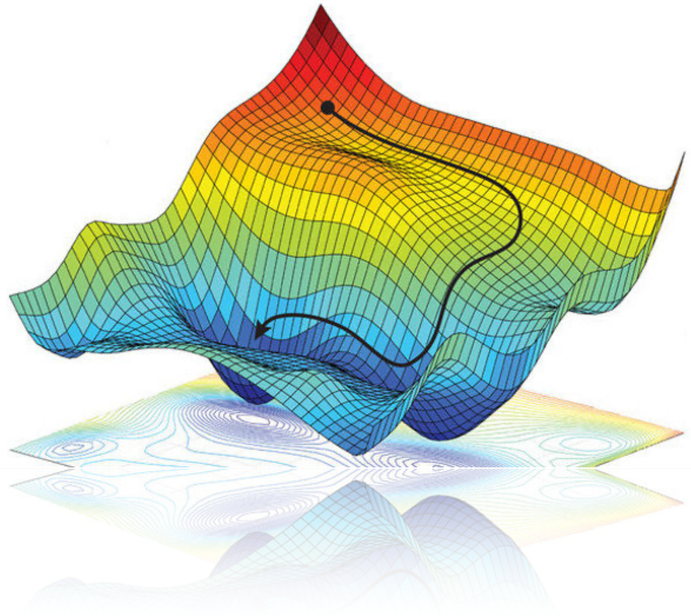
$$a_t \sim \pi_\theta(\cdot \mid s_t)$$

Como calcular o gradiente da nossa função objetivo em relação aos parâmetros da política?

$$\max_\theta J(\theta) = \max_\theta \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} [R(\tau)]$$

Com uma aproximação do gradiente, temos um método iterativo de melhorar a política

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$



Como calcular o gradiente de uma esperança?

$$\nabla_\theta \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[R(\tau)] \neq \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[\nabla_\theta R(\tau)]$$

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] &= \nabla_\theta \int P(\tau|\theta) R(\tau) d\tau \\
&= \int \nabla_\theta (P(\tau|\theta) R(\tau)) d\tau \\
&\neq \int P(\tau|\theta) \nabla_\theta R(\tau) d\tau = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta R(\tau)]
\end{aligned}$$

Do gradiente da esperança à esperança do gradiente (1/3)

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)]$$

$$= \nabla_\theta \int_\tau P(\tau \mid \theta) R(\tau) d\tau$$

(Definição de esperança)

Do gradiente da esperança à esperança do gradiente (2/3)

$$\mathbb{E}_{\tau \sim \pi_\theta} \left[\nabla_\theta \log P(\tau|\theta) R(\tau) \right]$$

Qual a probabilidade de uma trajetória ao seguirmos uma política?

Do gradiente da esperança à esperança do gradiente (3/3)

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\nabla_\theta \log P(\tau \mid \theta) R(\tau) \right]$$

$$= \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R(\tau) \right]$$

Note que:

- Não precisamos do gradiente da dinâmica do ambiente!
- Não é necessário que o retorno seja diferenciável

Calculando o gradiente de esperanças (caso geral)

Quando a distribuição de uma variável aleatória depende dos parâmetros $x \sim p(\cdot;\ \theta)$

E queremos calcular o gradiente da esperança de uma função objetivo $\frac{\partial}{\partial\theta}\mathbb{E}_x\left[f(x)\right]$

Podemos usar o ***score function estimator*** para aproximar o gradiente usando amostras

$$\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p_\theta}\left[f(x)\right] = \mathbb{E}_{x\sim p_\theta}\left[f(x)\frac{\partial}{\partial\theta}\log p_\theta(x)\right]$$

Obs.: o gradiente do log da probabilidade de uma v.a. é conhecido como seu *score (*https://en.wikipedia.org/wiki/Score_(statistics)*)*

Aproximando esperanças por Monte Carlo

- Em geral, o cálculo de um valor esperado é intratável
- Na prática, aproximamos a solução com uma média empírica

$$\mathbb{E}_{x\sim p}\left[f(x)\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

$$x_i \sim p(x_i)$$

Estimação por Monte Carlo é uma das principais ferramentas que possibilita Aprendizado por Reforço!

Aproximando esperanças por Monte Carlo

Aplicando esse conceito ao nosso *Policy Gradient*,

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R(\tau) \right]$$

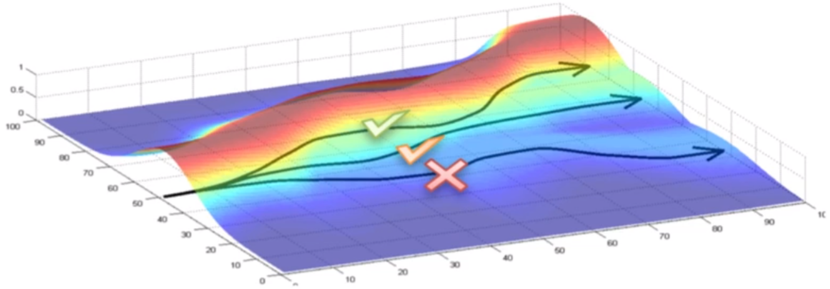
$$\approx \frac{1}{N} \sum_{k=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^k \mid s_t^k) R(\tau^k)$$

$$\tau^k = (s_0^k, a_0^k, r_1^k, s_1^k, \ldots, s_{T-1}^k, a_{T-1}^k, r_T^k, s_T^k) \sim \pi_\theta$$

Intuição: Policy Gradient

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R(\tau) \right]$$

- O score de uma trajetória indica a direção de aumento de sua probabilidade
- Trajetórias de retorno positivo tem sua probabilidade aumentada
- O contrário acontece quando o retorno é negativo



$R(\tau) > 0$	$\uparrow \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)$	**Reforço positivo**
$R(\tau) < 0$	$\downarrow \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)$	**Reforço negativo**

Sumário da abordagem

$$R^k = \sum_{t=0}^{T-1} r_{t+1}^k$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{k=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^k \mid s_t^k) R^k$$

Estimação de → **Atualização**

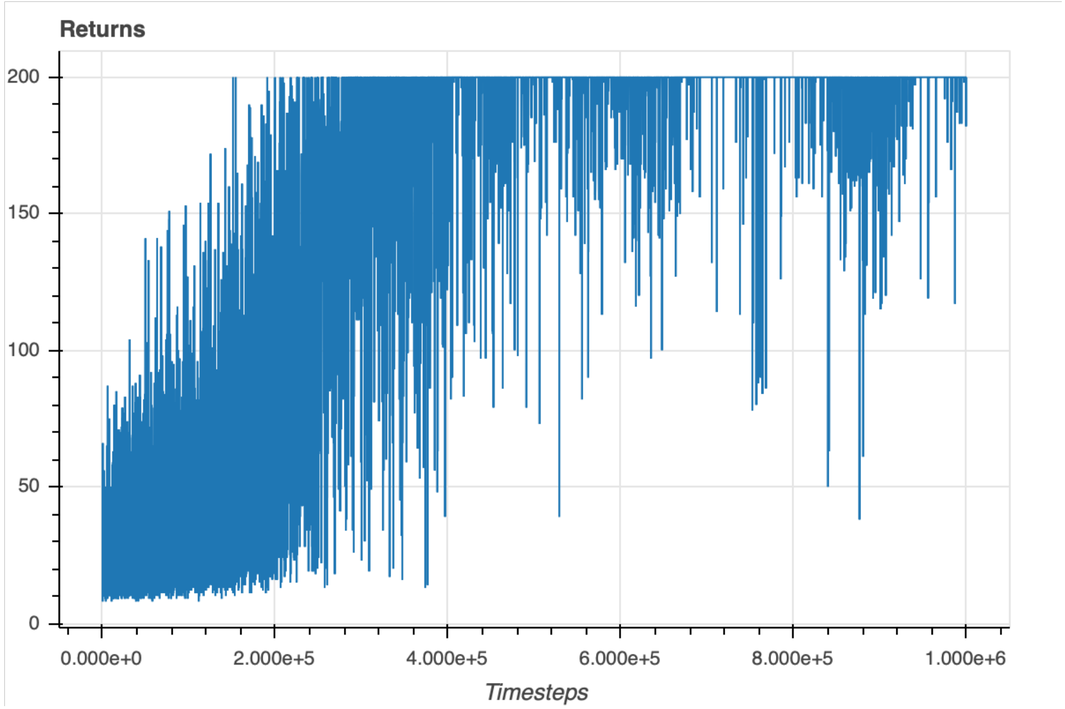
Interação

$$\tau_1, \tau_2, \ldots, \tau_N \sim \pi_\theta$$

Desafios e limitações

$$\mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R(\tau) \right]$$

- O policy gradient depende da distribuição sob a política atual
- Após uma atualização, devemos coletar novos dados
- O gradiente tende a ser ruidoso (alta variância)



REINFORCE: combinando Policy Gradients e Monte Carlo

Algoritmo 1 REINFORCE

Entrada: parâmetros da política, θ

1: **enquanto** não satisfeito **faça**

2: Colete trajetórias com a política atual, $\tau_1, \tau_2, \ldots, \tau_N \sim \pi_\theta$

3: Calcule os retornos de cada trajetória, $R_k = \sum_{t=1}^{T} r_t^k$

4: Estime o *Policy Gradient*

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{k=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^k \mid s_t^k) R_k$$

5: Atualize os parâmetros da política, $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

6: **fim enquanto**

7: **devolve** π_θ

Referências

(1) OpenAI SpinningUp Part 3: Intro to Policy Optimization

- https://spinningup.openai.com/en/latest/spinningup/rl_intro3.html

(2) Sutton & Barto; Reinforcement Learning: An Introduction (2nd Edition)

- Chapter 13: Policy Gradient Methods

(3) An overview of gradient descent optimization algorithms

- http://arxiv.org/abs/1609.04747