



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

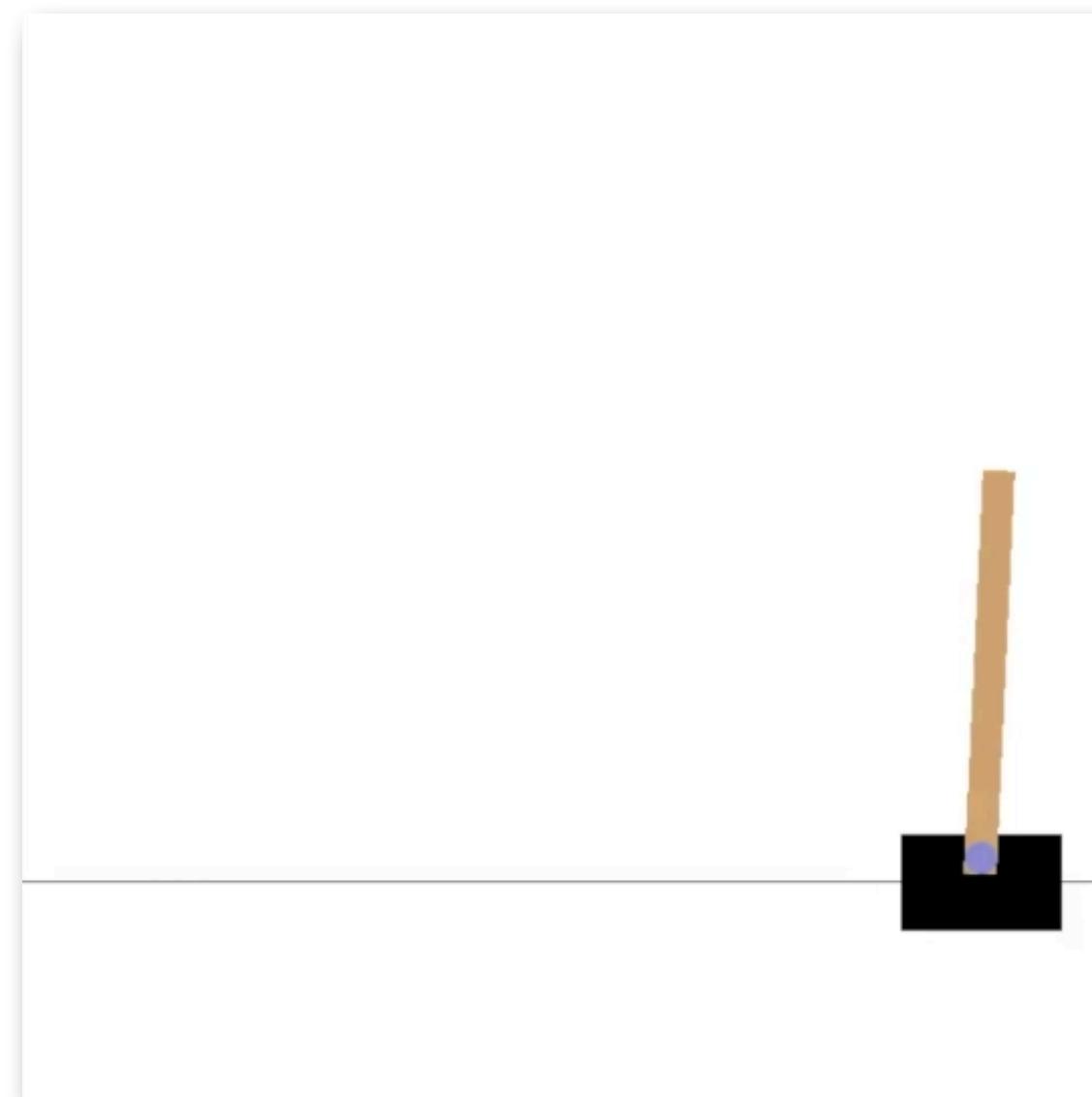
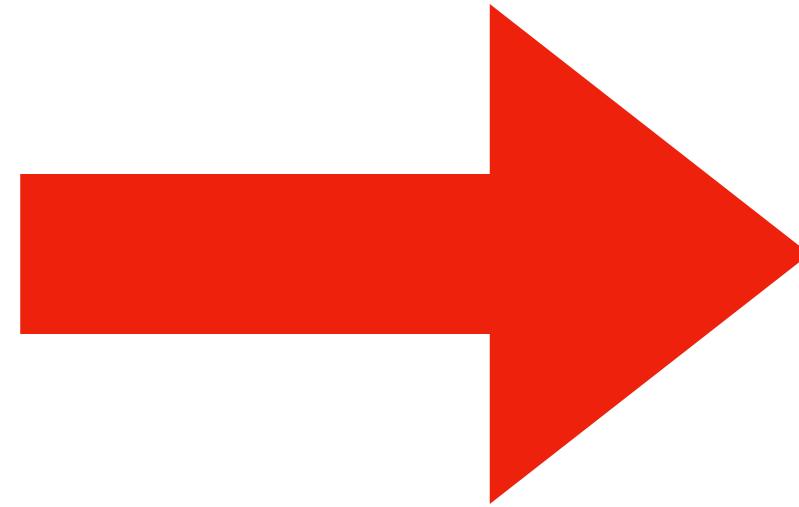
1. Introdução
2. Derivação do gradiente de política
3. Algoritmo (REINFORCE)

Objetivos

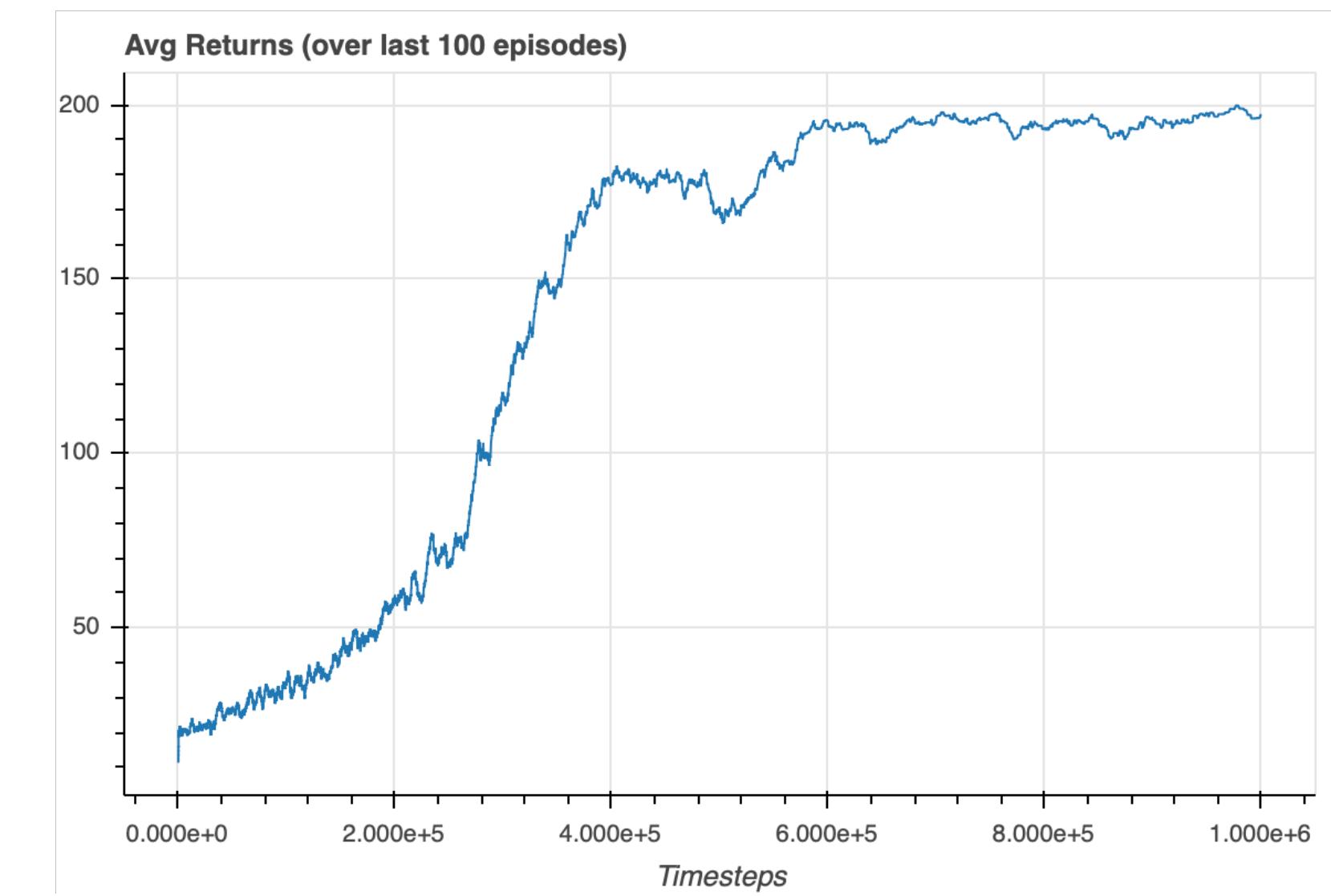
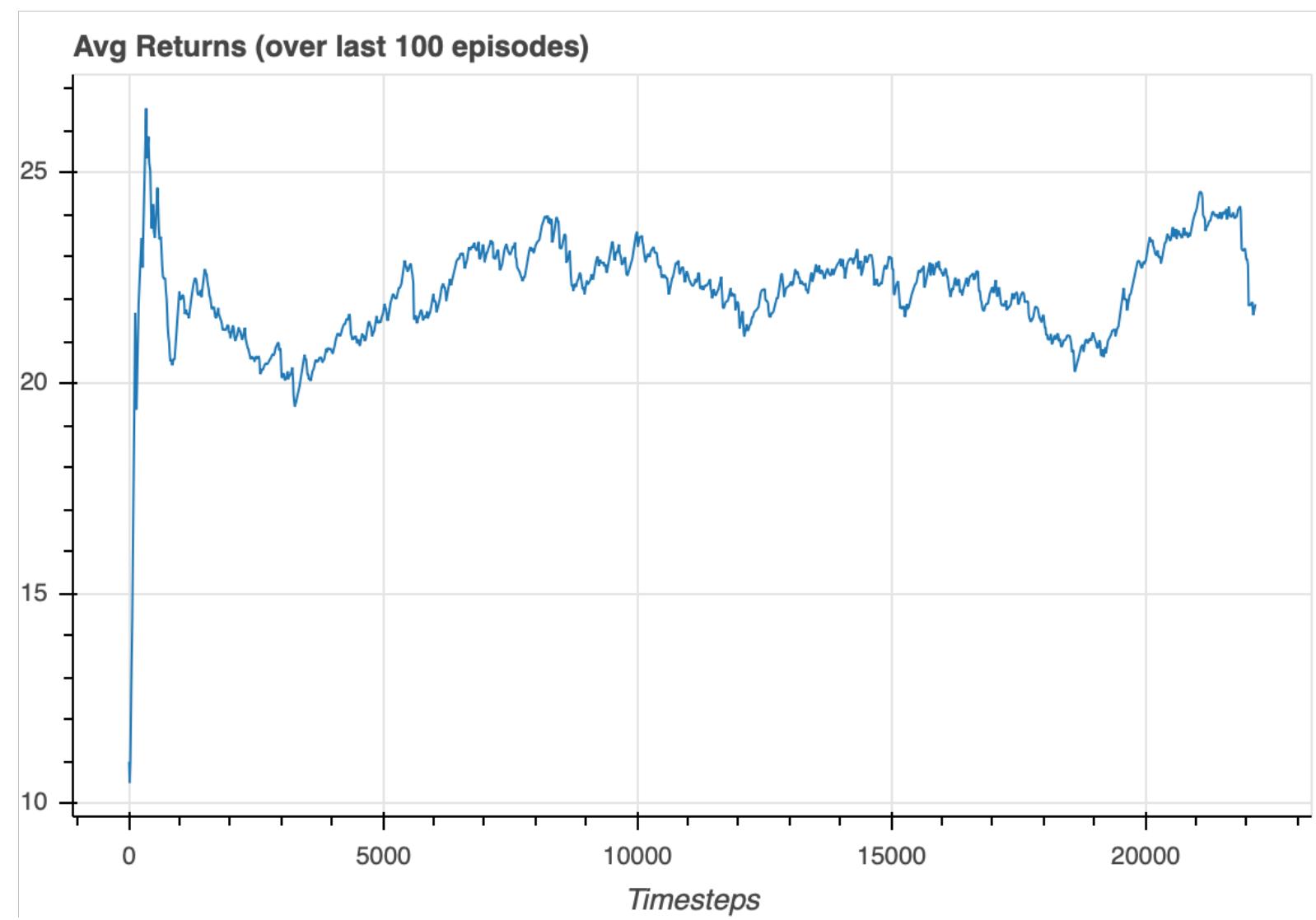
- Entender (model-free) RL como abordagem trial-and-error
- Entender abordagem de Policy Gradients como busca no espaço de parâmetros da política
- Implementar algoritmo REINFORCE via diferenciação automática



Aula de hoje

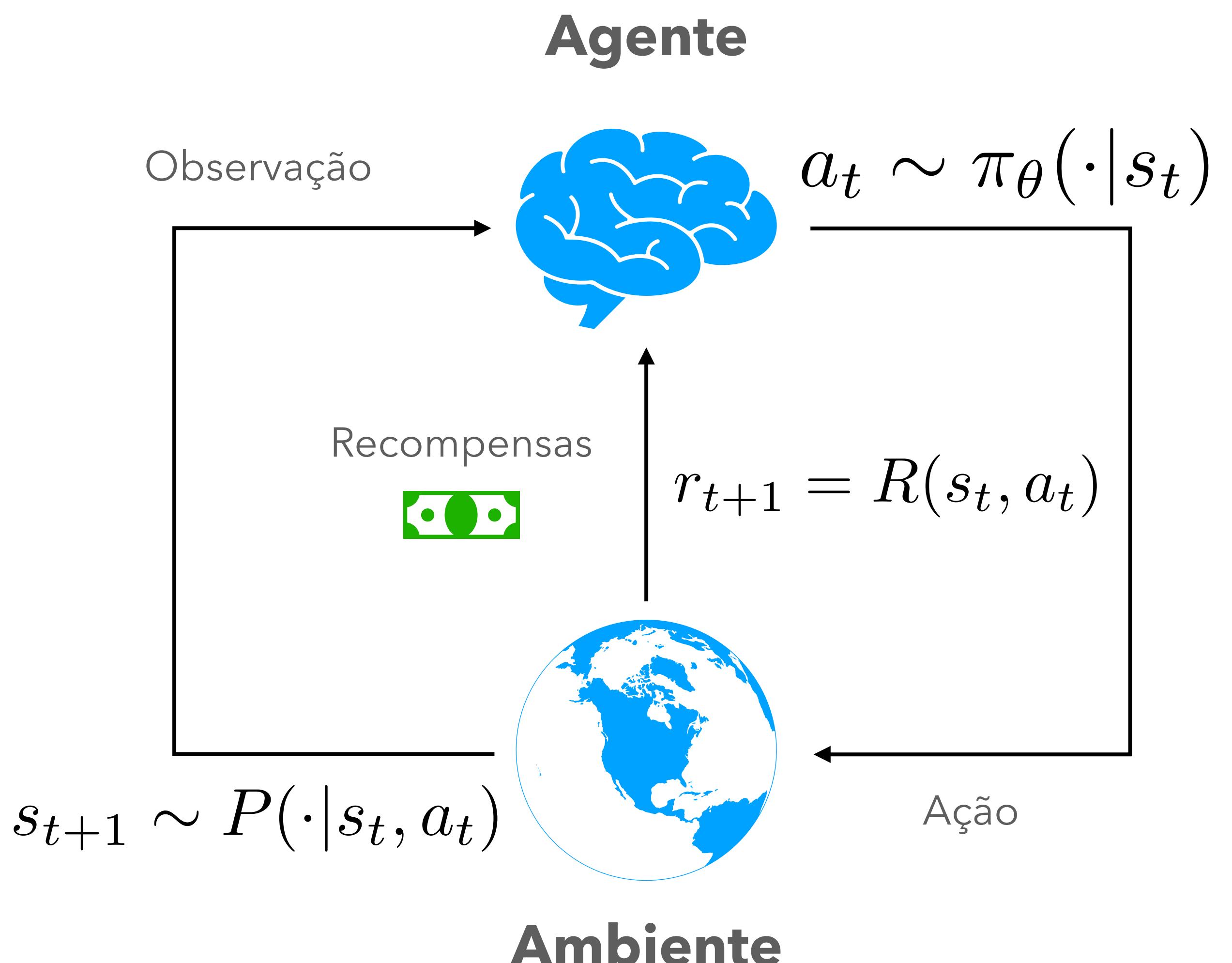


Policy Gradient



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



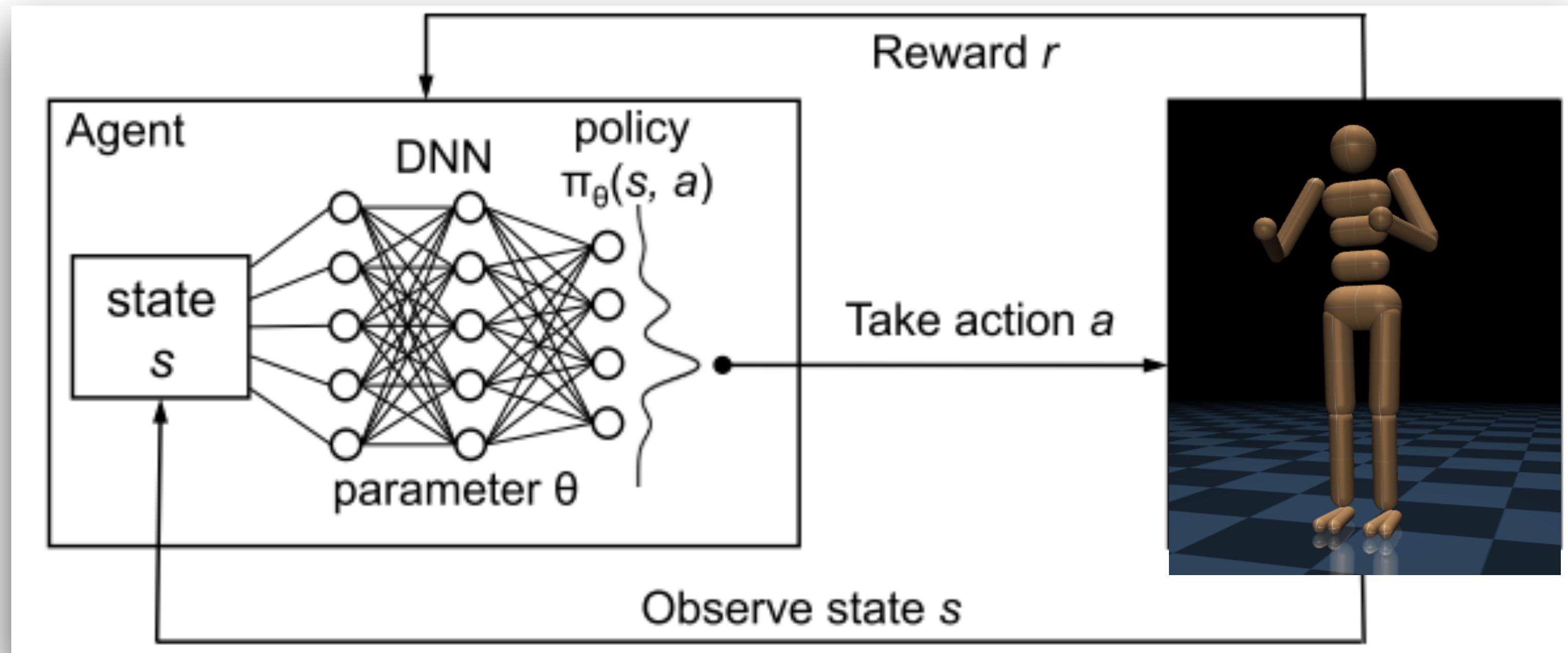
$$\tau = (s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T, s_T)$$

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

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

- 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

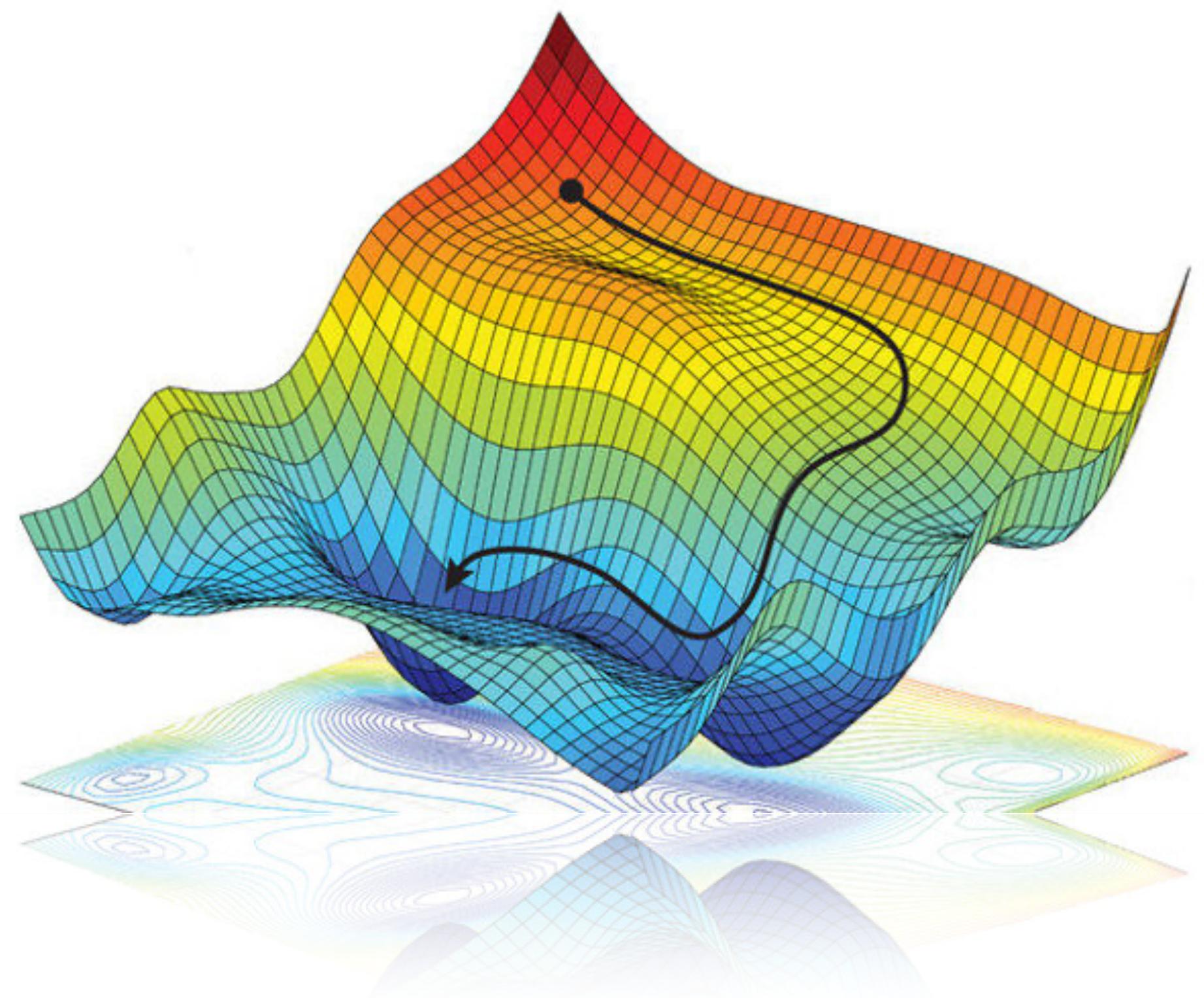
$$a_t \sim \pi_\theta(\cdot | s_t)$$

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

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

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?

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

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



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

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)] \\ &= \nabla_{\theta} \int_{\tau} P(\tau \mid \theta) R(\tau) d\tau \quad (\text{Definição de esperança})\end{aligned}$$



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

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)] \\&= \nabla_{\theta} \int_{\tau} P(\tau \mid \theta) R(\tau) d\tau && \text{(Definição de esperança)} \\&= \int_{\tau} \nabla_{\theta} P(\tau \mid \theta) R(\tau) d\tau && \text{(Regra de Leibniz)}\end{aligned}$$



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

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)] \\&= \nabla_{\theta} \int_{\tau} P(\tau \mid \theta) R(\tau) d\tau && \text{(Definição de esperança)} \\&= \int_{\tau} \nabla_{\theta} P(\tau \mid \theta) R(\tau) d\tau && \text{(Regra de Leibniz)} \\&= \int_{\tau} P(\tau \mid \theta) \nabla_{\theta} \log P(\tau \mid \theta) R(\tau) d\tau && (\nabla \log f(x) = \nabla f(x)/f(x))\end{aligned}$$



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

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)] \\&= \nabla_{\theta} \int_{\tau} P(\tau \mid \theta) R(\tau) d\tau && (\text{Definição de esperança}) \\&= \int_{\tau} \nabla_{\theta} P(\tau \mid \theta) R(\tau) d\tau && (\text{Regra de Leibniz}) \\&= \int_{\tau} P(\tau \mid \theta) \nabla_{\theta} \log P(\tau \mid \theta) R(\tau) d\tau && (\nabla \log f(x) = \nabla f(x)/f(x)) \\&= \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau \mid \theta) R(\tau)] && (\text{Definição de esperança})\end{aligned}$$



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

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

Qual a probabilidade de uma trajetória ao seguirmos uma política?



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

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

Qual a probabilidade de uma trajetória ao seguirmos uma política?

$$P(\tau | \theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t | s_t) P(s_{t+1} | s_t, a_t)$$



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

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

Qual a probabilidade de uma trajetória ao seguirmos uma política?

$$P(\tau | \theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t | s_t) P(s_{t+1} | s_t, a_t)$$

$$\log P(\tau | \theta) = \log \rho_0(s_0) + \sum_{t=0}^{T-1} \log \pi_\theta(a_t | s_t) + \log P(s_{t+1} | s_t, a_t)$$



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

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

Qual a probabilidade de uma trajetória ao seguirmos uma política?

$$P(\tau | \theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t | s_t) P(s_{t+1} | s_t, a_t)$$

$$\log P(\tau | \theta) = \log \rho_0(s_0) + \sum_{t=0}^{T-1} \log \pi_\theta(a_t | s_t) + \log P(s_{t+1} | s_t, a_t)$$

Como os parâmetros da política influenciam essa probabilidade?

$$\nabla_\theta \log P(\tau | \theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t)$$



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

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau | \theta) R(\tau)] \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]\end{aligned}$$

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

$$\frac{\partial}{\partial \theta} \mathbb{E}_x [f(x)]$$

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

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

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

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

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*,

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

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

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

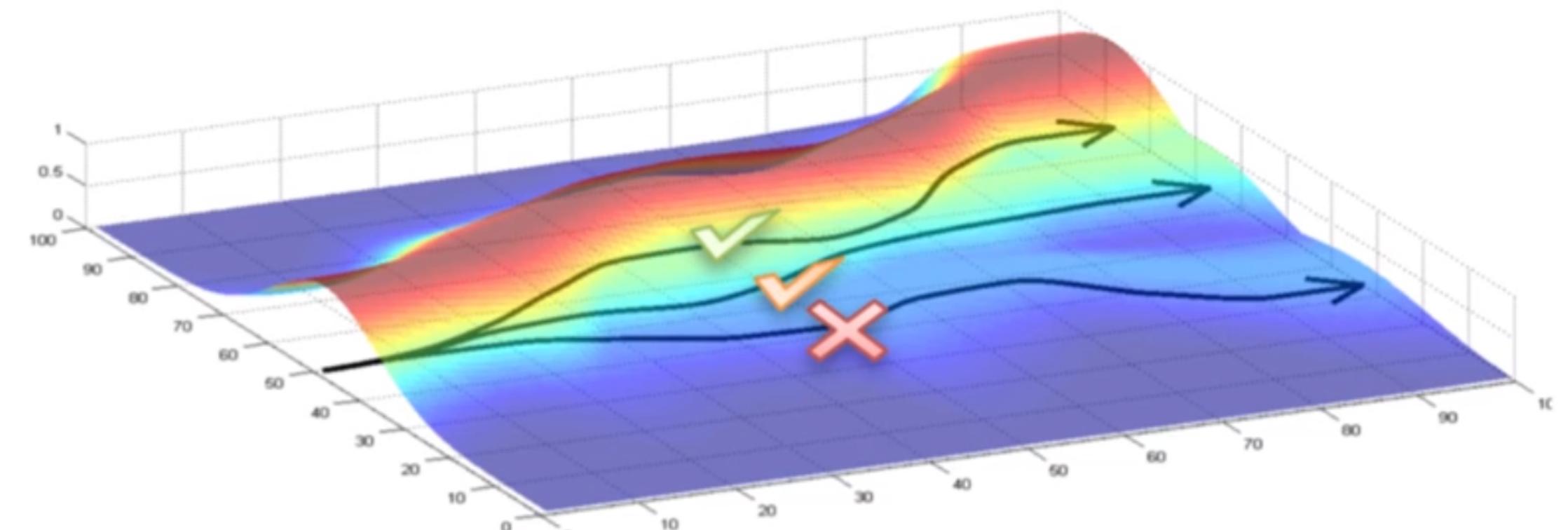
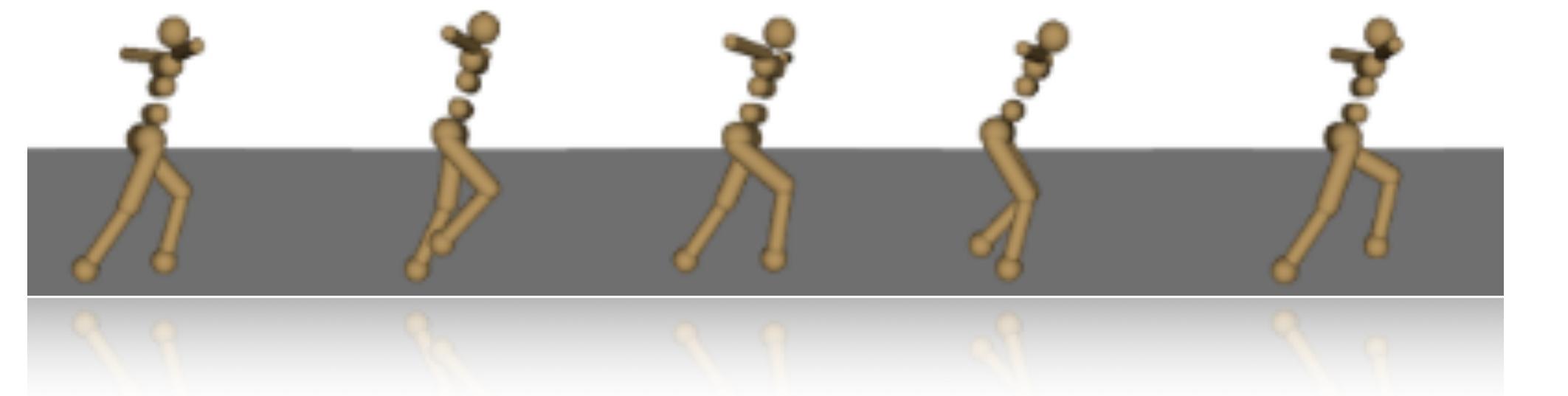


Intuição: Policy Gradient

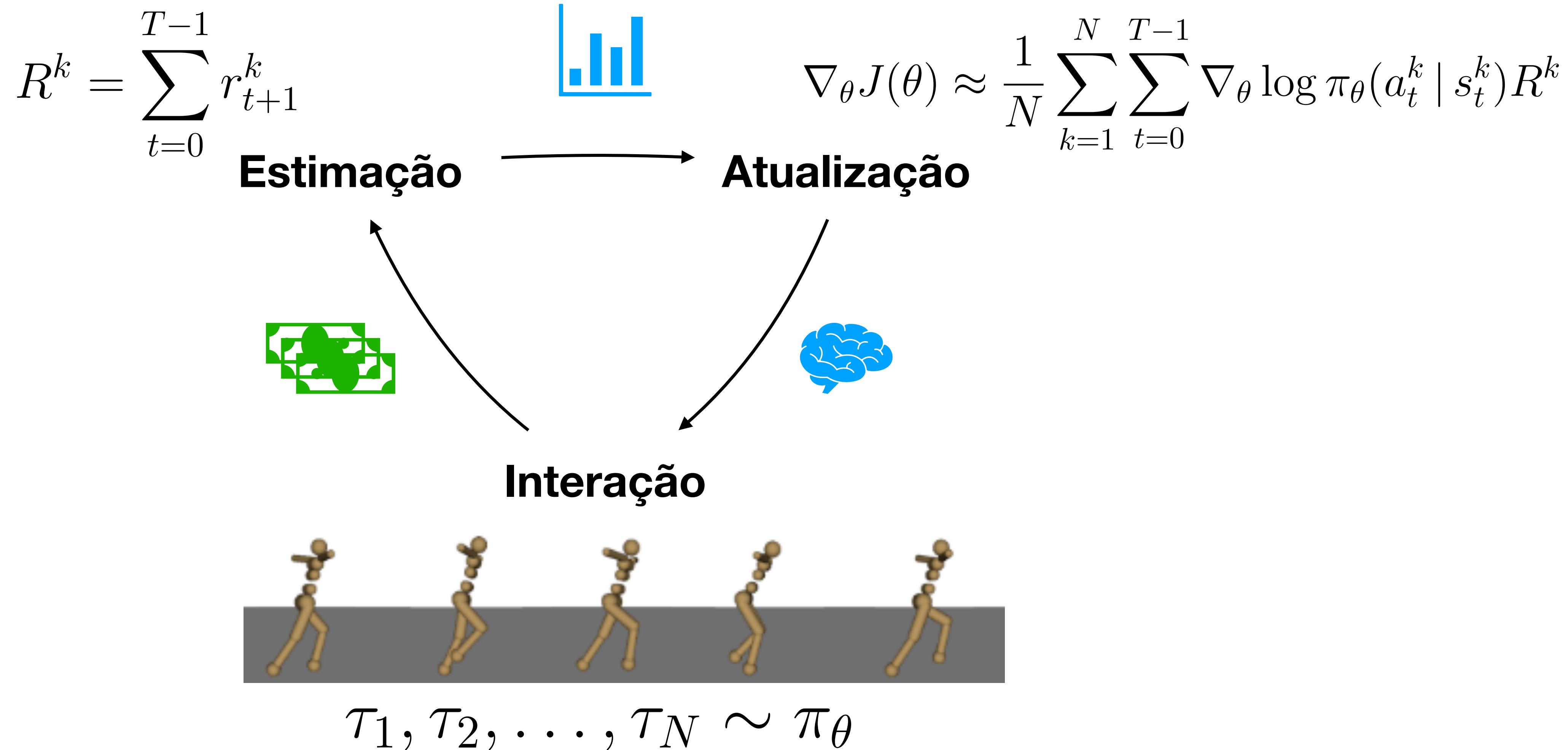
$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]$$

- 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

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



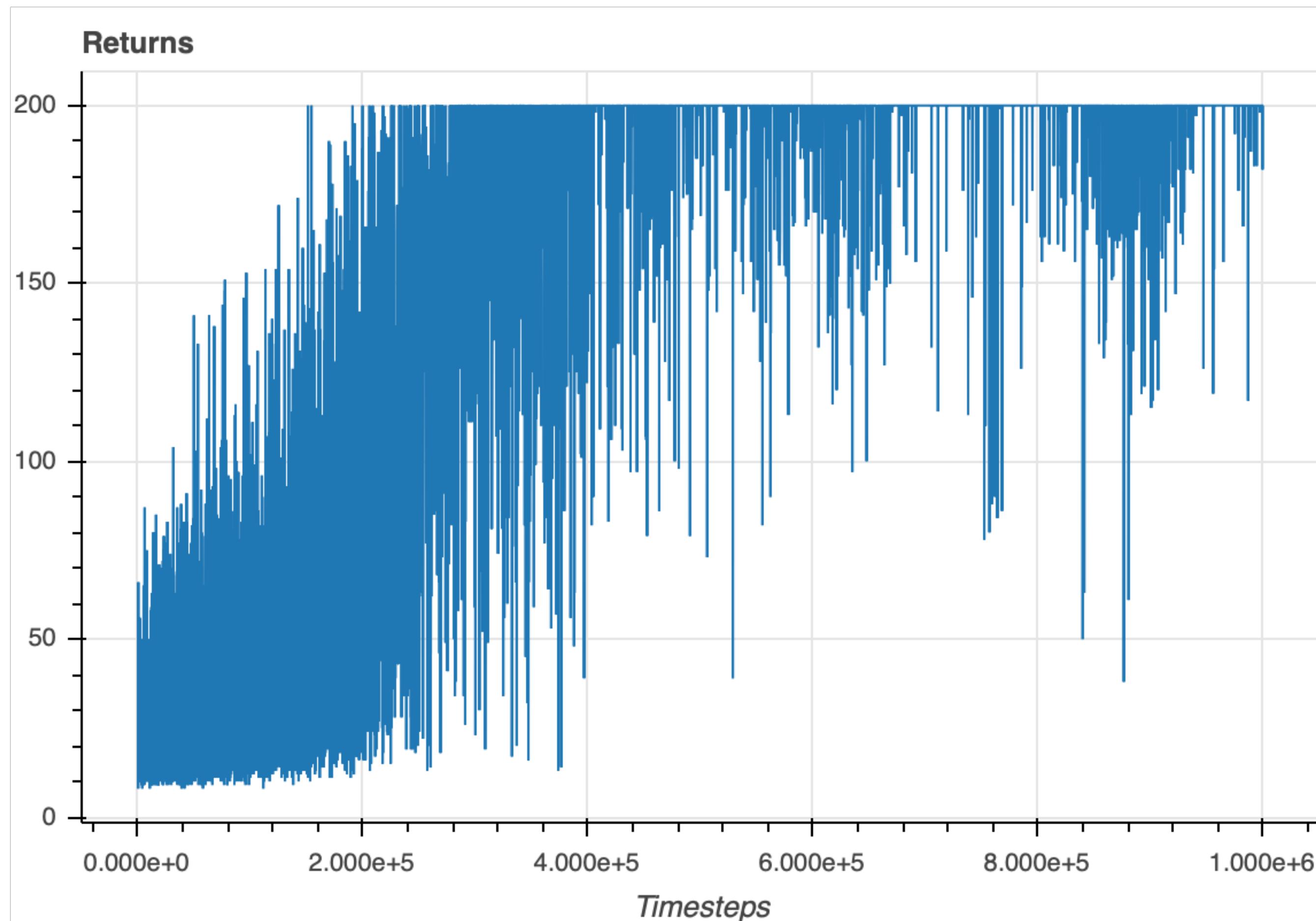
Sumário da abordagem



Desafios e limitações

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

- 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, $\tau_1, \tau_2, \dots, \tau_N \sim \pi_\theta$
- 3: Calcule os retornos de cada trajetória, $R_k = \sum_{t=1}^T r_t^k$
- 4: Estime o *Policy Gradient*

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

- 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>

