

# **APRENDIZAGEM POR REFORÇO**

Luís Morgado

ISEL-ADEETC

# EXPLORAÇÃO vs. APROVEITAMENTO

- Estimativas de valor de uma acção  $a$

- $Q_t(a) \approx Q^*(a)$

- Acção de aproveitamento (*greedy*)

- $a_t^* = \operatorname{argmax}_a [Q_t(a)]$

- Aproveitamento

- $a_t = a_t^*$

- Exploração

- $a_t \neq a_t^*$

# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

- Estratégia *greedy*

$$a_t = a_t^* = \operatorname{argmax}_a Q_t(a)$$

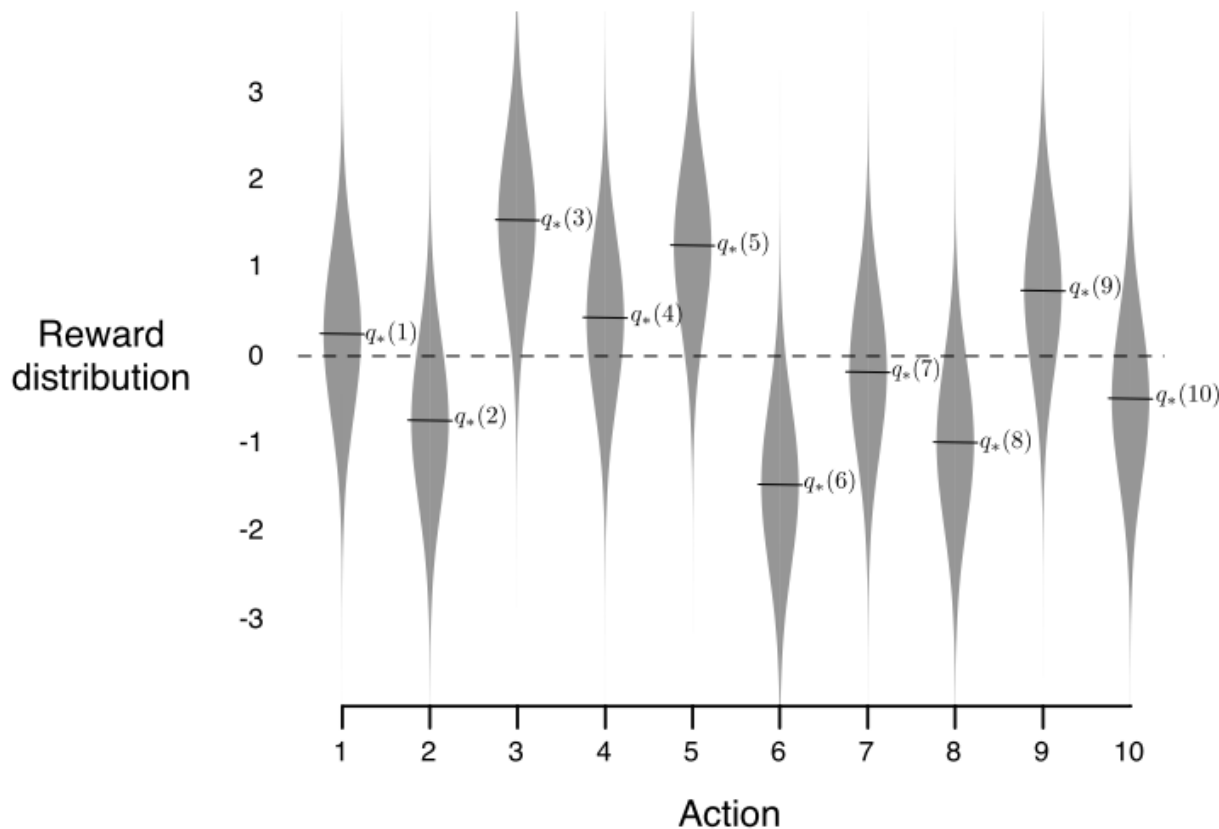
- Estratégia  $\varepsilon$ -*greedy*

$$a_t = \begin{cases} a_t^* & \text{com probabilidade } 1 - \varepsilon \\ \text{acção aleatória} & \text{com probabilidade } \varepsilon \end{cases}$$

– Balanceamento de *exploração* vs. *aproveitamento*

## 2.3 The 10-armed Testbed

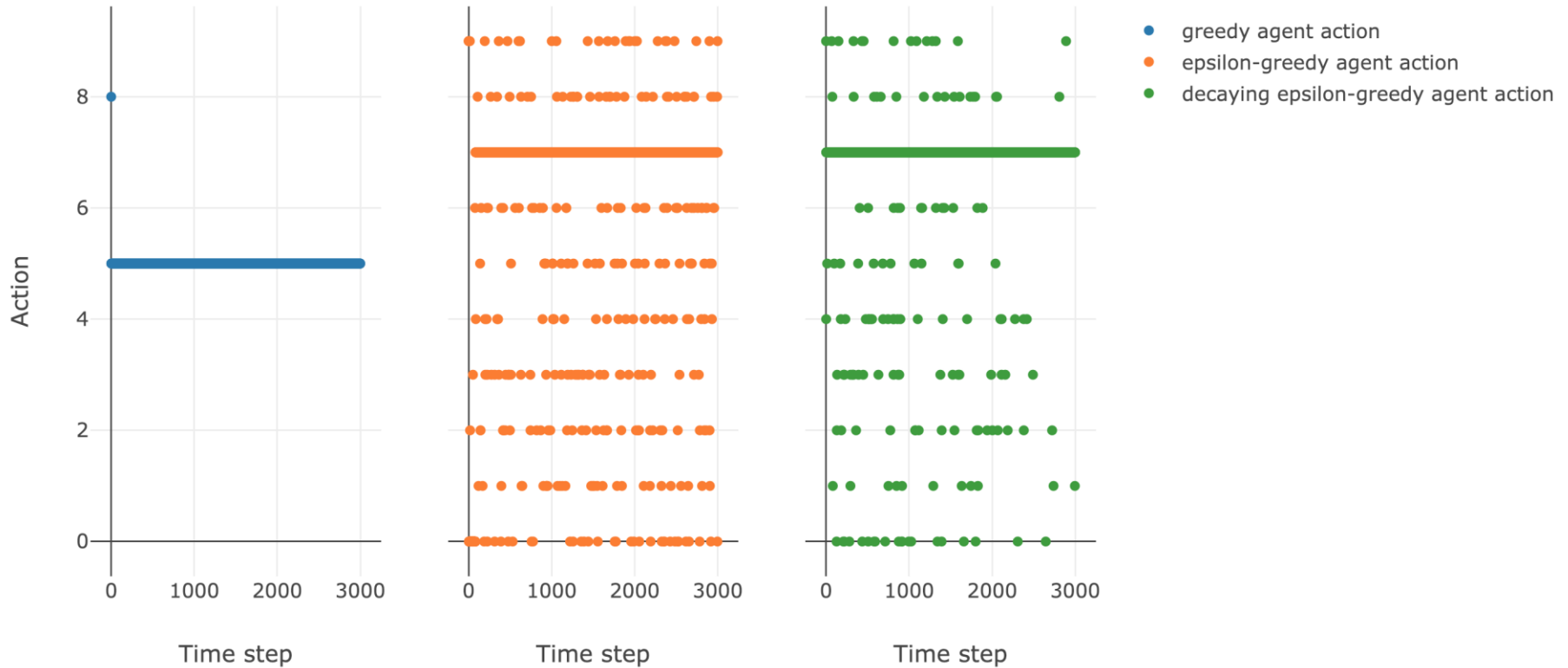
To roughly assess the relative effectiveness of the greedy and  $\varepsilon$ -greedy action-value methods, we compared them numerically on a suite of test problems. This was a set of 2000 randomly generated  $k$ -armed bandit problems with  $k = 10$ . For each bandit problem, such as the one shown in Figure 2.1, the action values,  $q_*(a)$ ,  $a = 1, \dots, 10$ ,



**Figure 2.1:** An example bandit problem from the 10-armed testbed. The true value  $q_*(a)$  of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean  $q_*(a)$ , unit-variance normal distribution, as suggested by these gray distributions.

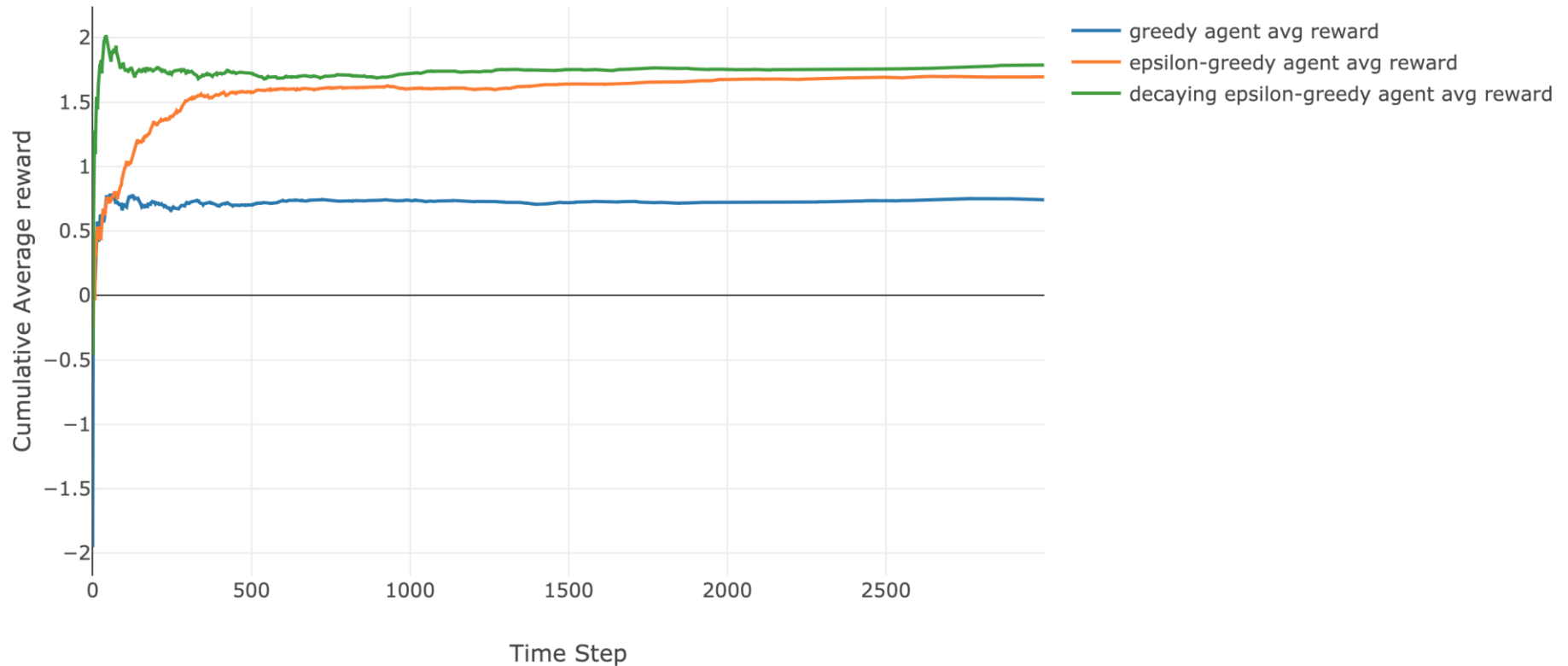
# Exemplo

Actions taken by different agents



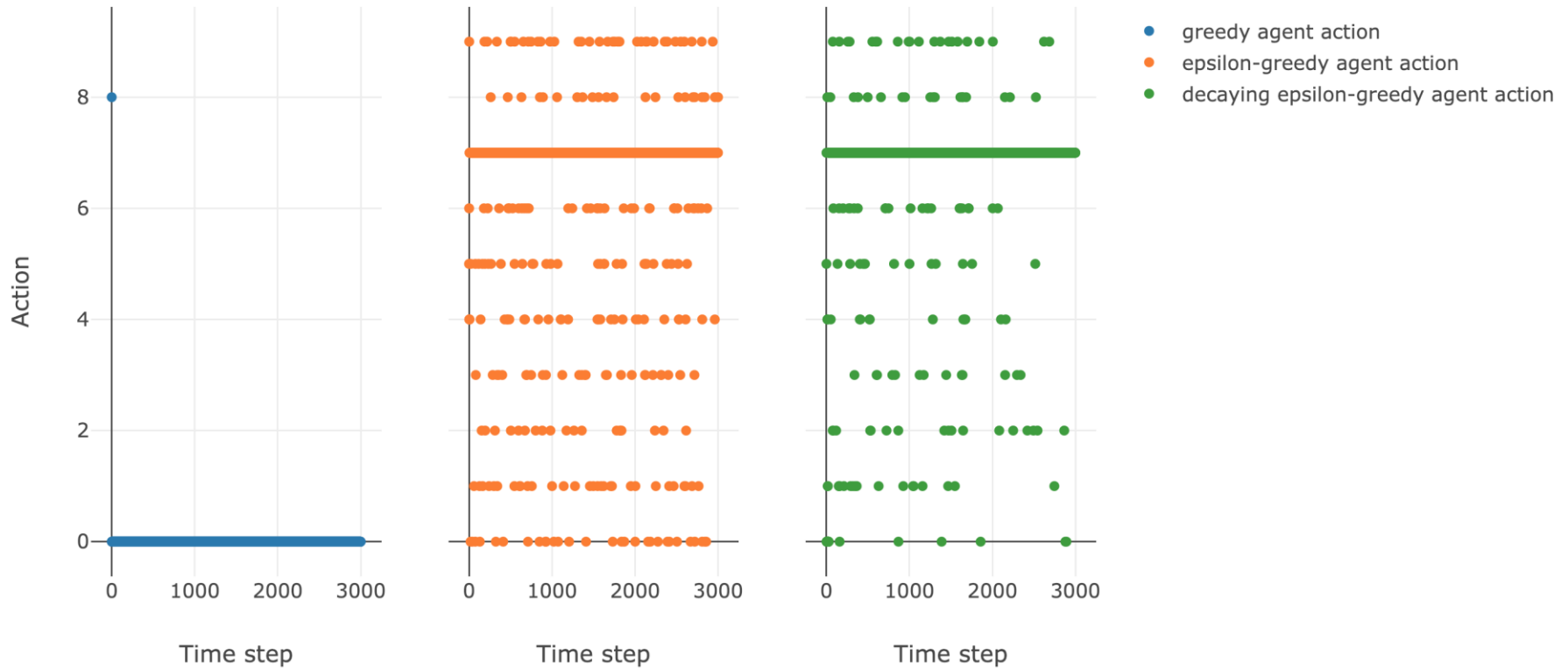
# Exemplo

Cumulative Average reward recieved over time



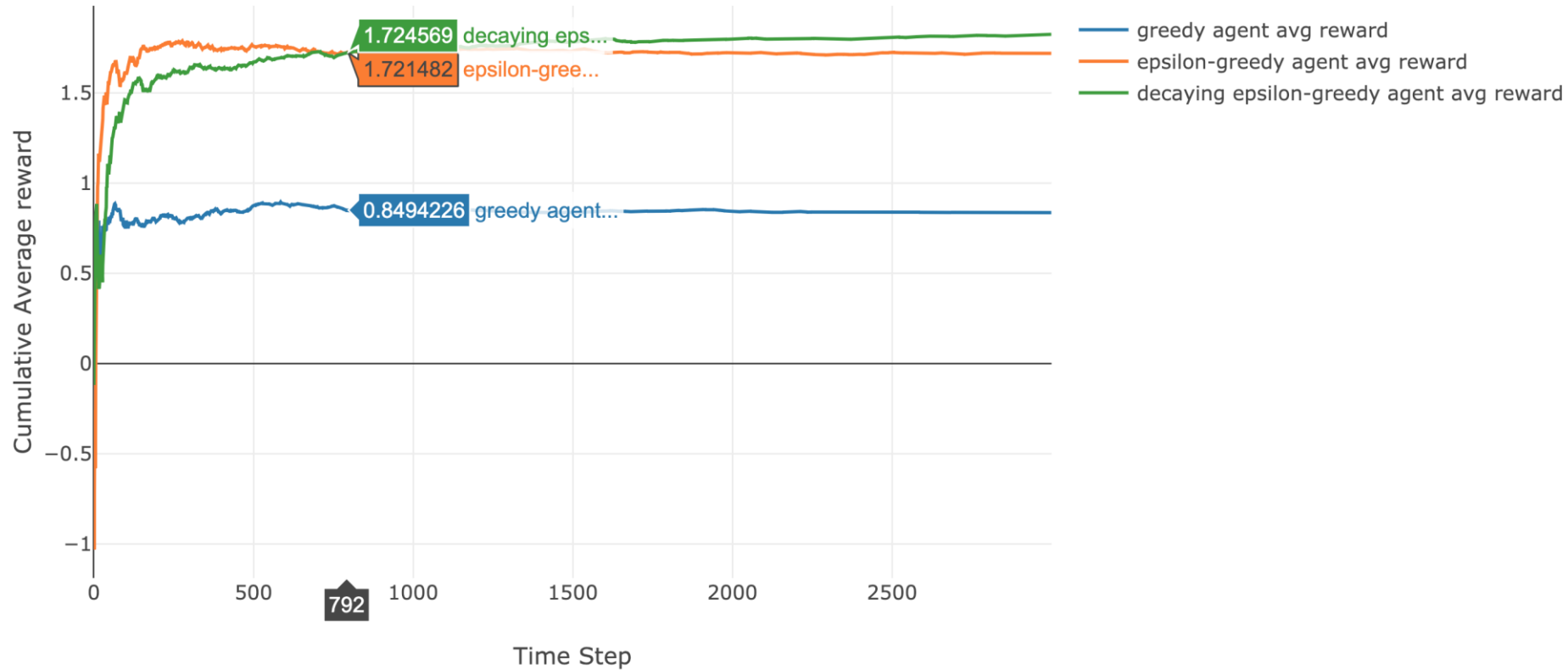
# Exemplo

Actions taken by different agents



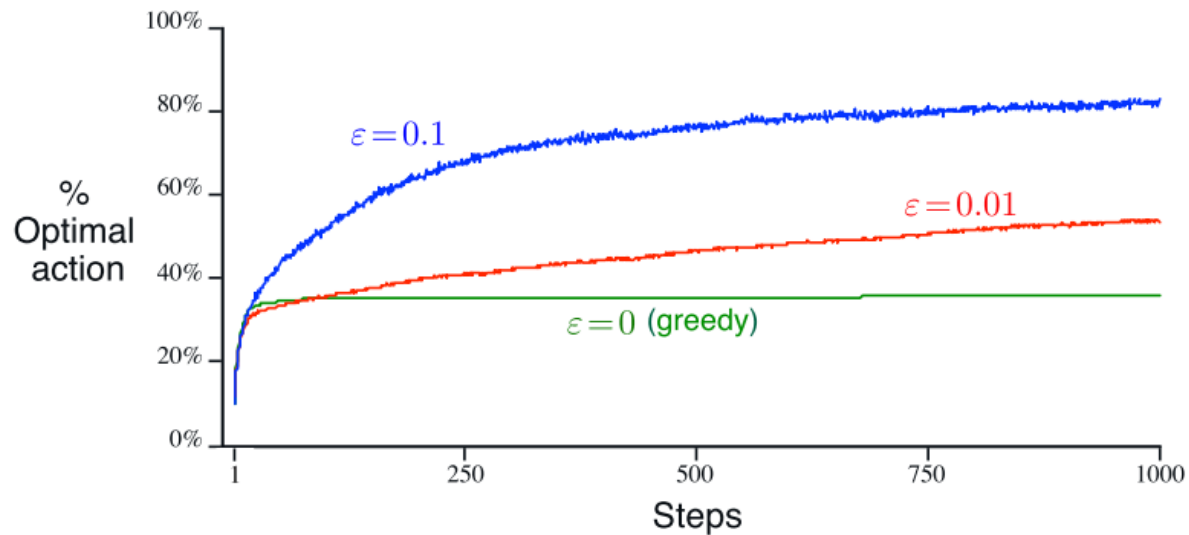
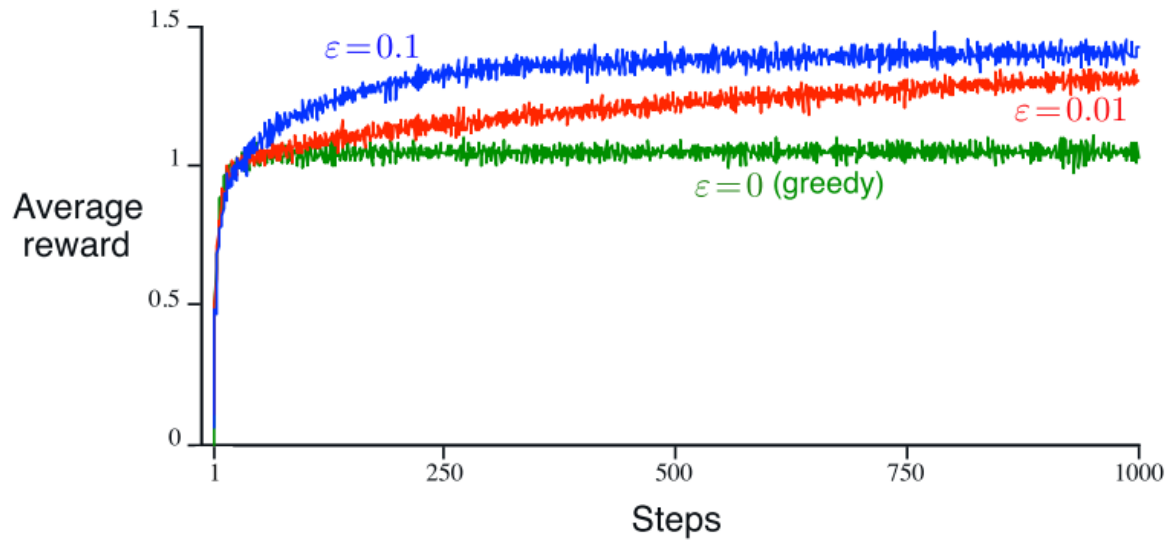
# Exemplo

Cumulative Average reward recieved over time



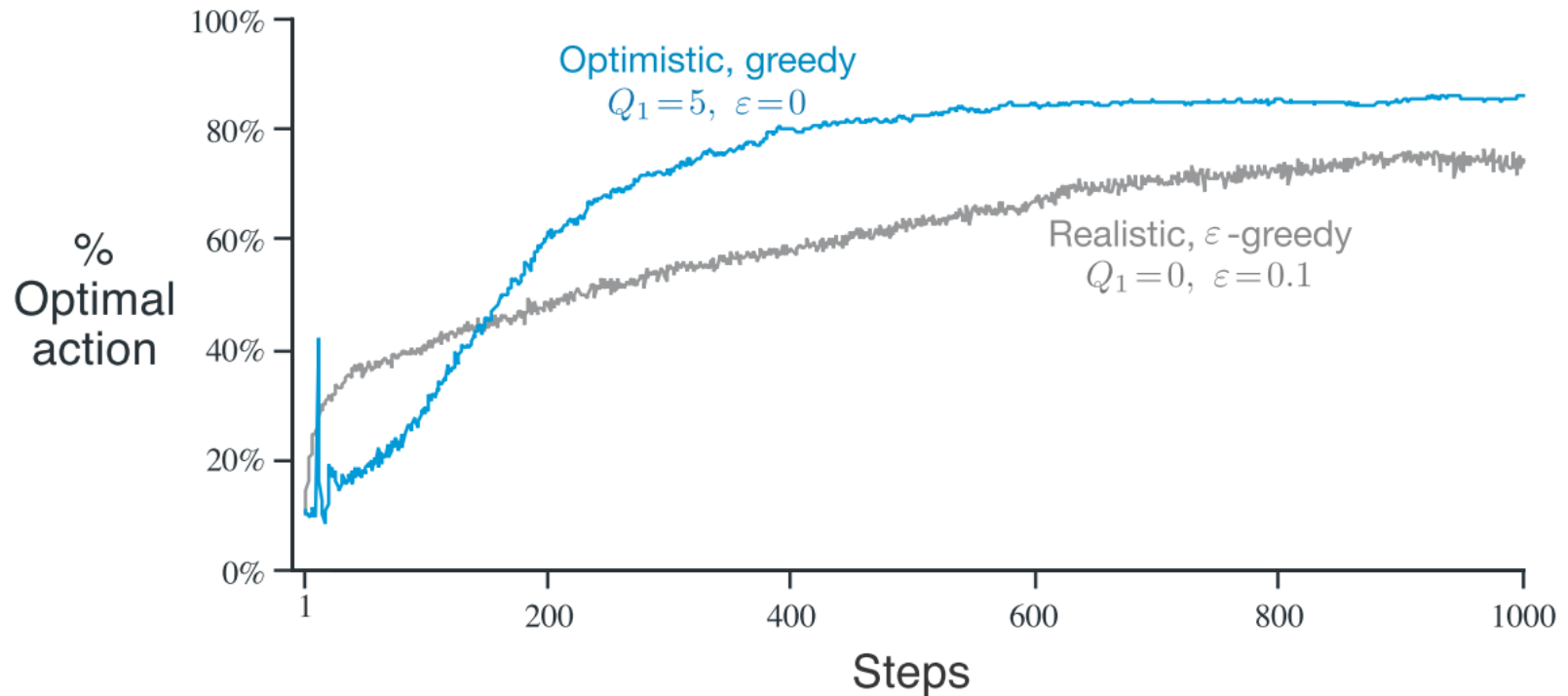


# Exemplo



# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

## Valores iniciais optimistas



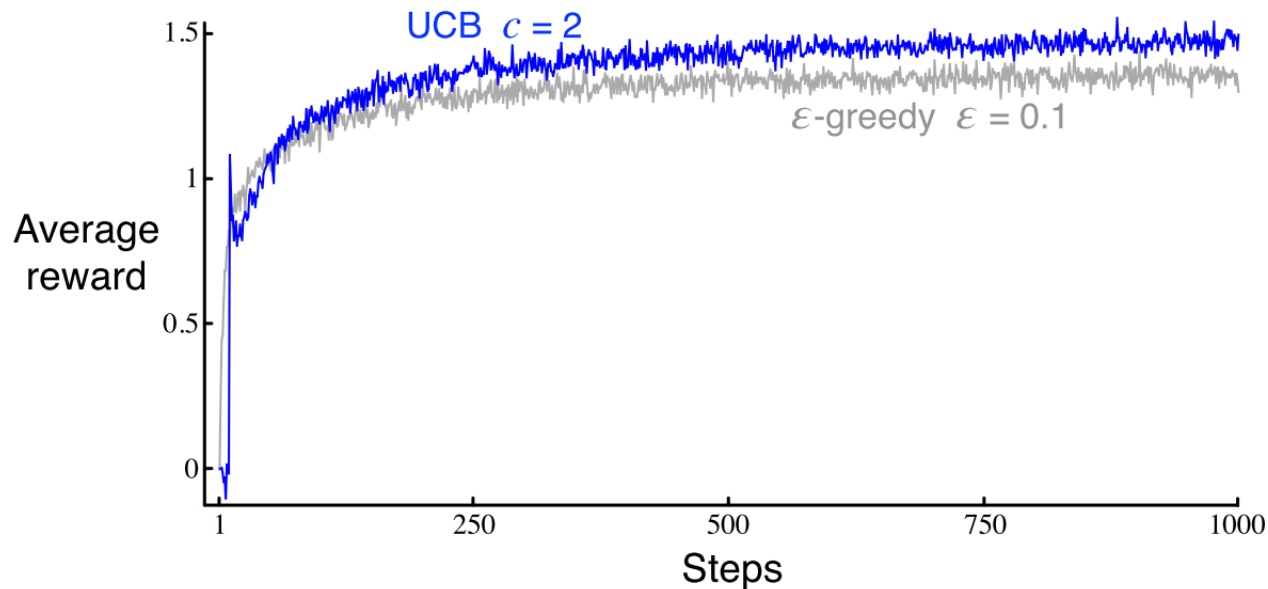
# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

## *Upper Confidence Bounds (UCB)*

$$A_t \doteq \operatorname{argmax}_a \left[ Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

$N_t(a)$  denotes the number of times that action  $a$  has been selected prior to time  $t$

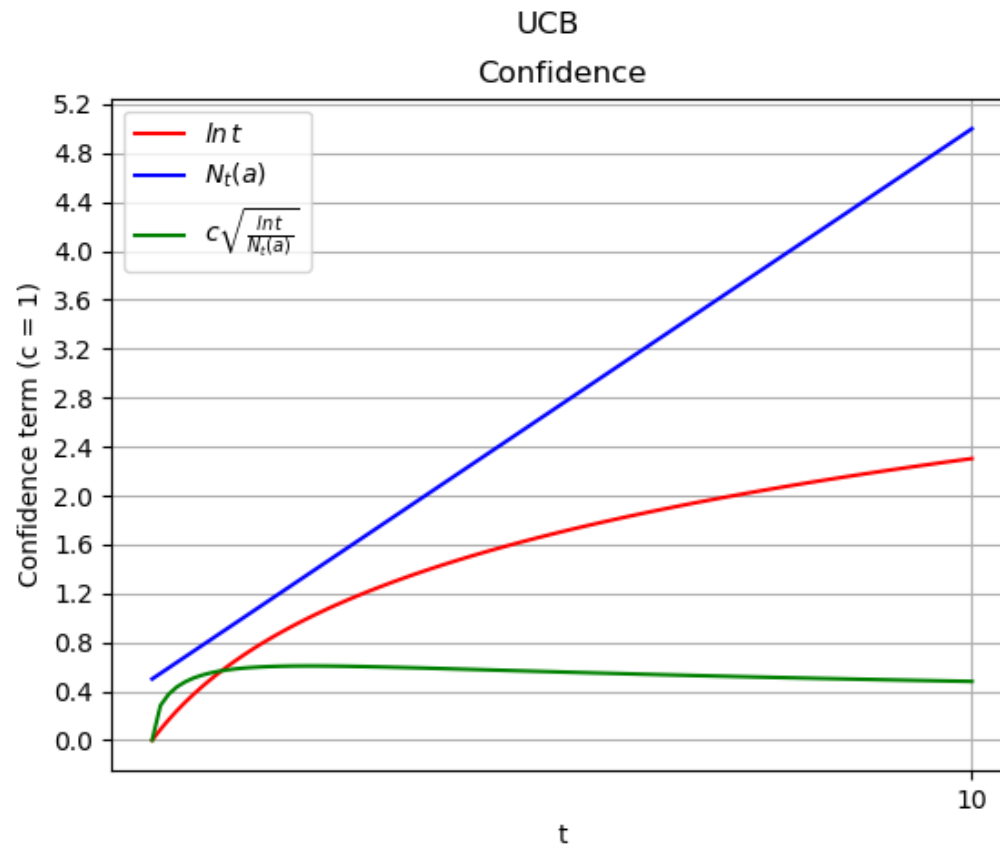
If  $N_t(a) = 0$ , then  $a$  is considered to be a maximizing action



# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

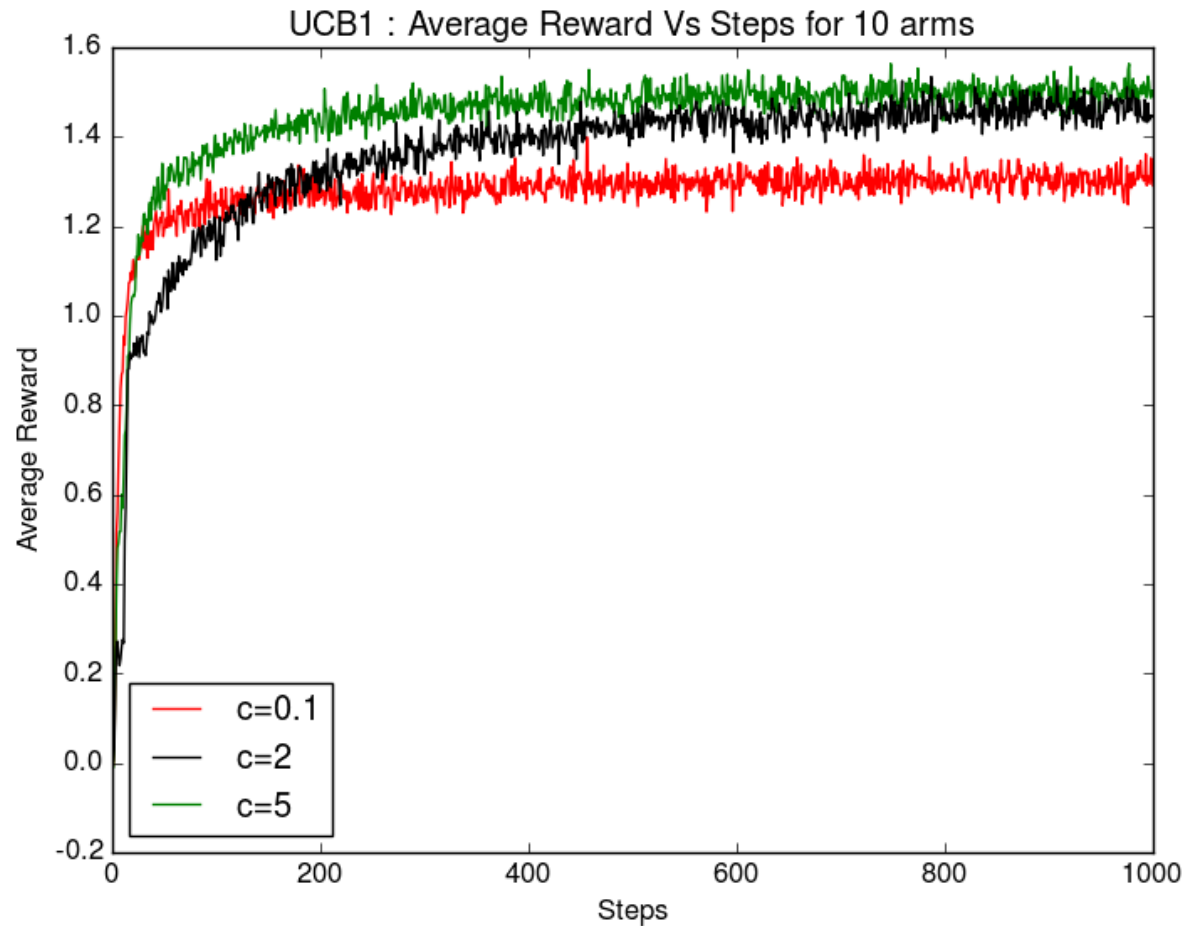
## *Upper Confidence Bounds (UCB)*

$$A_t \doteq \arg\max_a \left[ Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$



# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

## *Upper Confidence Bounds (UCB)*



# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

## Distribuição *Soft-max*

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^k e^{H_t(b)}}$$

$H_t(a) \in \mathbb{R}$  numerical *preference* for each action  $a$

$$H_t(a) = Q_t(a)/\tau$$

$$\Pr\{A_t = a\} = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^k e^{Q_t(b)/\tau}}$$

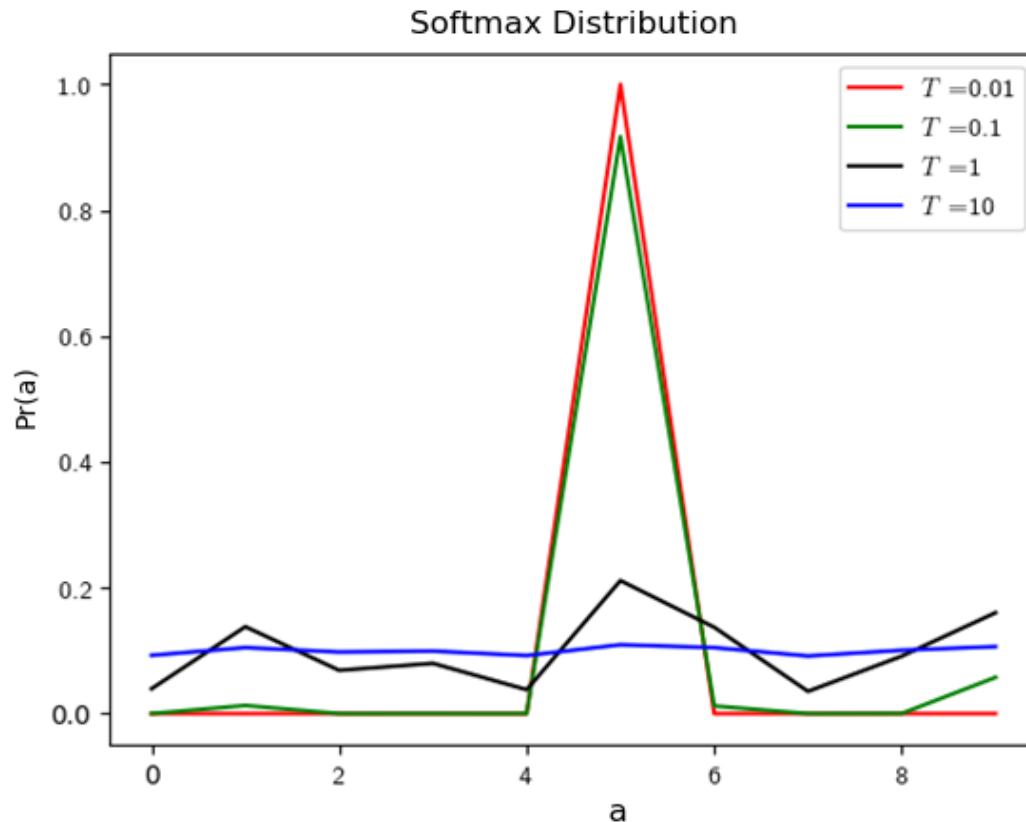
$\tau$  is a positive parameter called the *temperature*

In the limit as  $\tau \rightarrow 0$  softmax action selection becomes the same as greedy action selection

# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

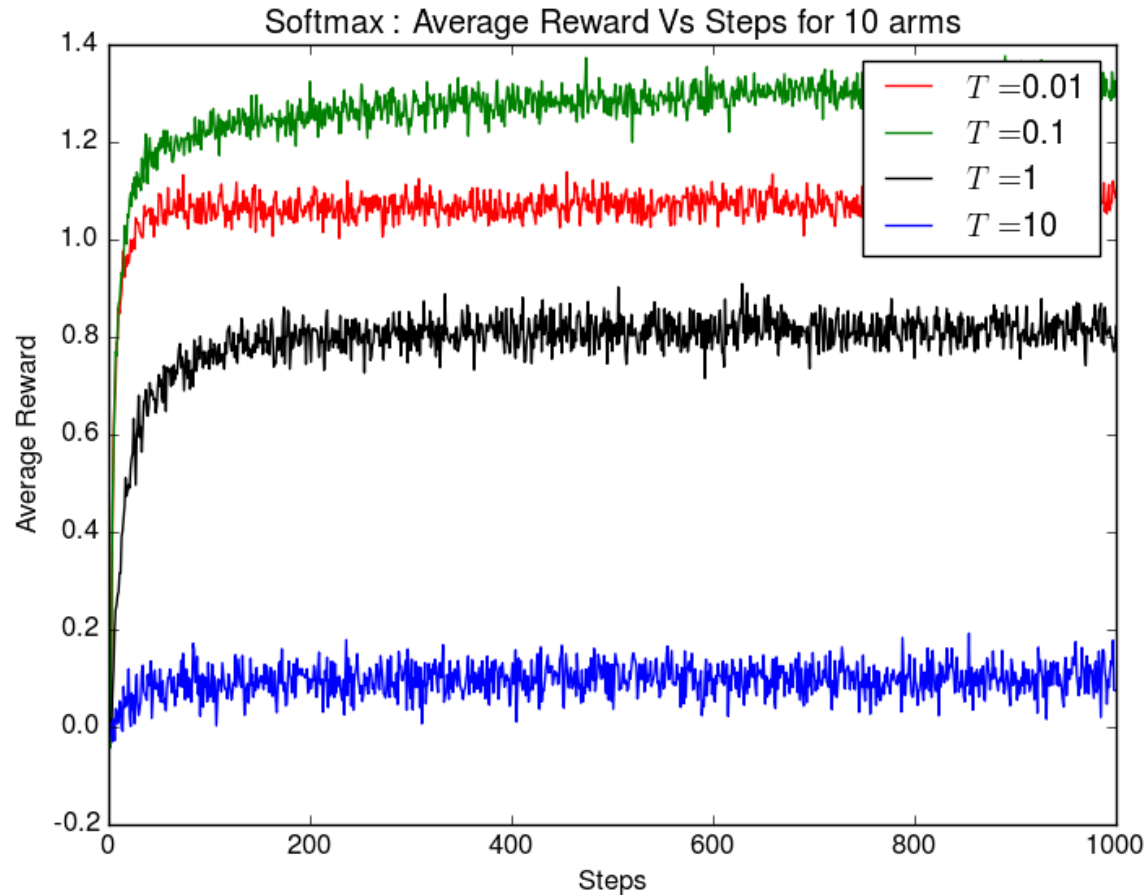
## Distribuição *Soft-max*

$$\Pr\{A_t = a\} = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^k e^{Q_t(b)/\tau}}$$



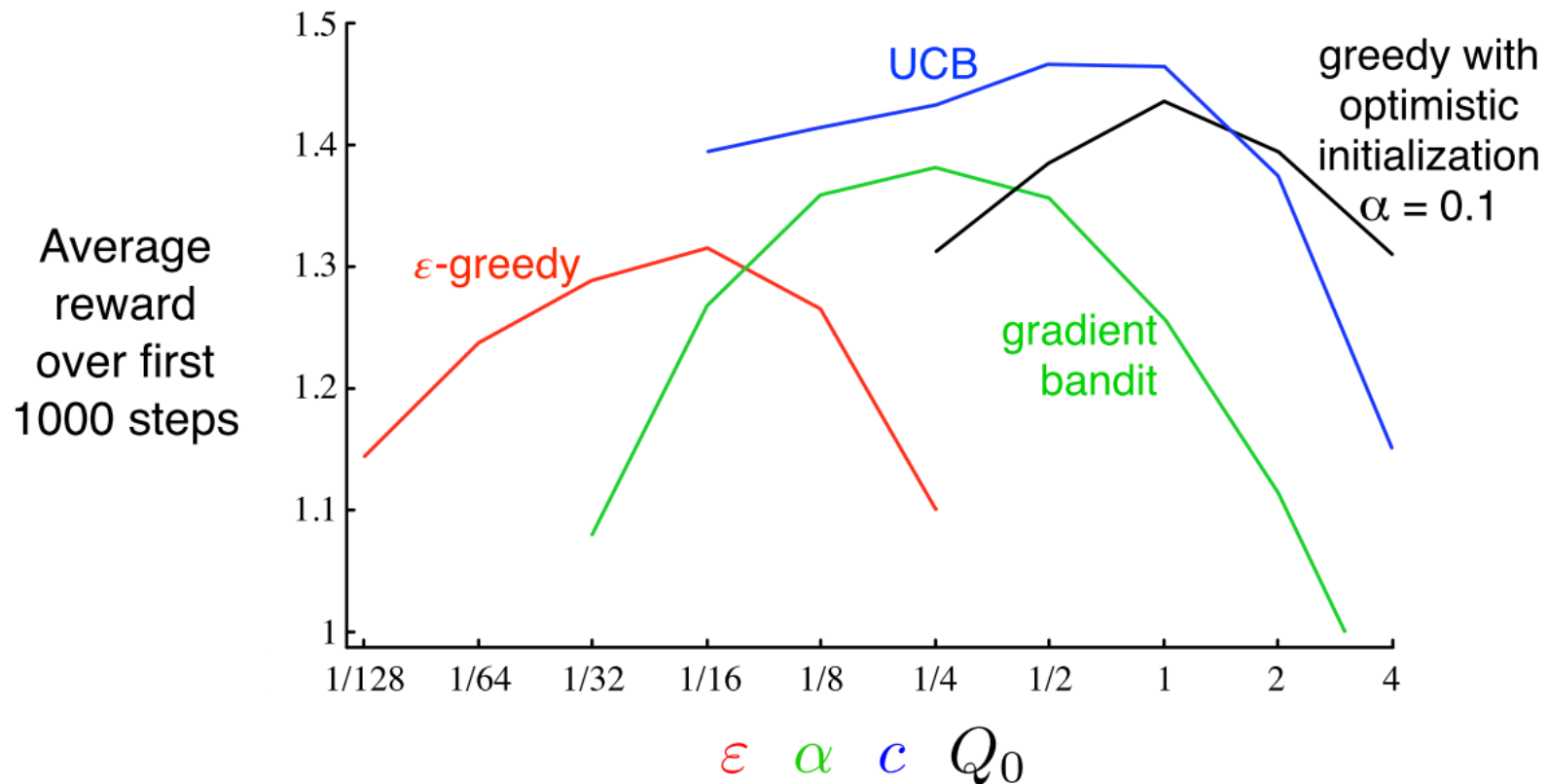
# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO

## Distribuição *Soft-max*





# ESTRATÉGIAS DE SELECÇÃO DE ACÇÃO



# REFERÊNCIAS

[Sutton & Barto, 2020]

R. Sutton, A. Barto, “Reinforcement Learning: An Introduction”, 2<sup>nd</sup> Edition, MIT Press, 2020

[Poole & Mackworth, 2010]

D. Poole, A. Mackworth, Artificial Intelligence: Foundations of Computational Agents, Cambridge University Press, 2010

[Barnard, 2003]

C. Barnard, “Animal Behaviour: Mechanism, Development, Ecology and Evolution”, Prentice Hall, 2003

[Koppula, 2020]

R. Koppula, “Exploration vs. Exploitation In Reinforcement Learning”, *<https://www.manifold.ai/exploration-vs-exploitation-in-reinforcement-learning>*, 2020