ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Cơ sở trí tuệ nhân tạo

Học tăng cường

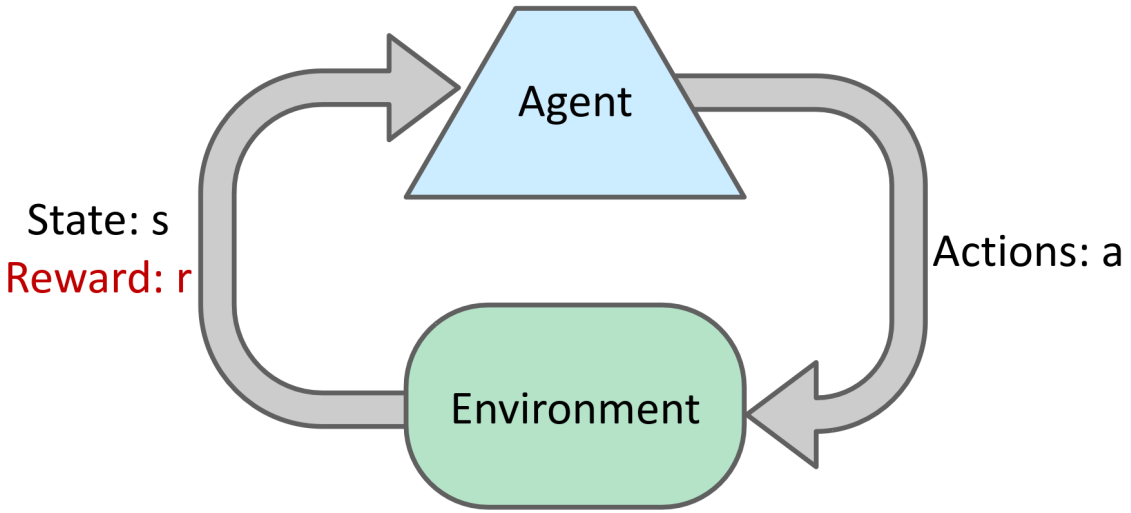
Nguyễn Ngọc Đức

2024

Nội dung

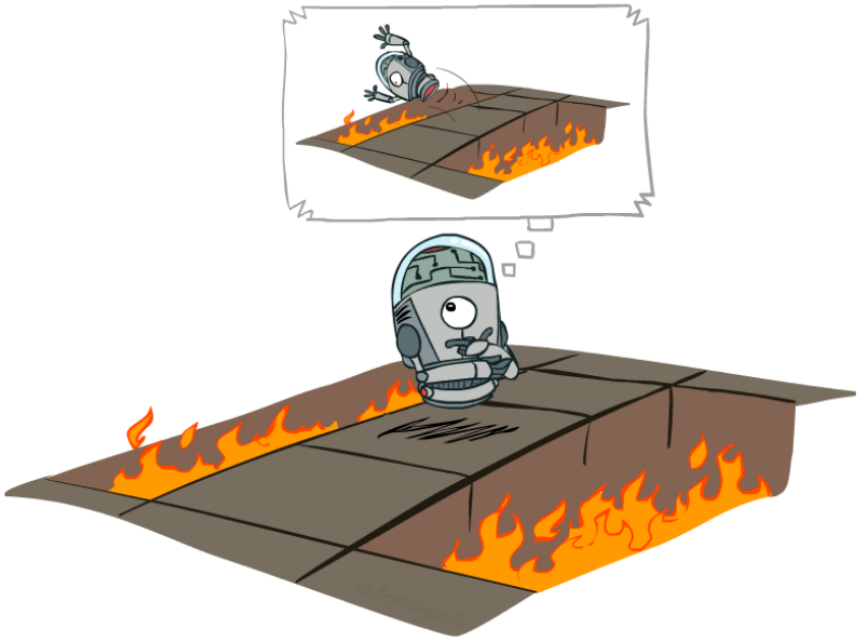
Học tăng cường

Học tăng cường I



- Một quy trình Markov với
  1. Hàm chuyển dịch $T(s, a, s')$ bất định
  2. Hàm điểm thưởng $R(s, a, s')$ bất định

Học tăng cường II



- Không biết trước successor của một trạng thái
- Không biết trước đánh giá của trạng thái
- **Thử và sai!!!**

Temporal Difference Learning

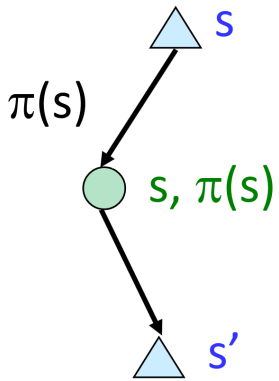
Temporal Difference Learning

- Ý tưởng chính: dựa trên kinh nghiệm!
  - Cập nhật $V(s)$ với mỗi hành động thực hiện
- Học hàm $V(s)$ theo thời gian:
  - Lấy mẫu kết quả (bằng cách thực hiện hành động)
  - Tính kỳ vọng

$$sample = R(s, \pi(s), s') + \gamma V^{\pi}(s')$$

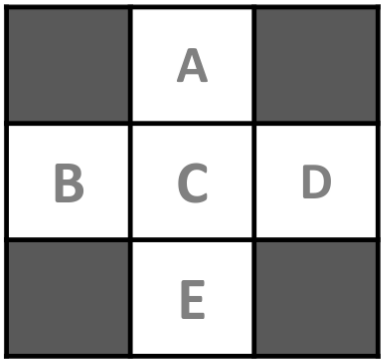
$$V^{\pi}(s) \leftarrow (1 - \alpha)V^{\pi}(s) + \alpha(sample)$$

$$V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha(sample - V^{\pi}(s))$$



Ví dụ

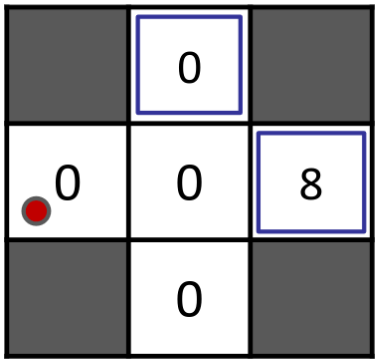
States



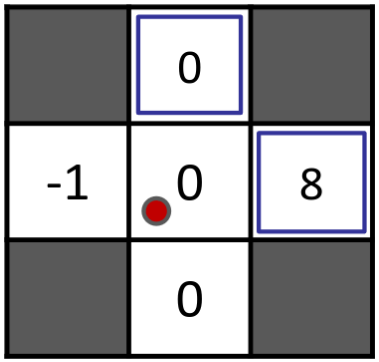
Assume: $\gamma = 1$, $\alpha = 1/2$

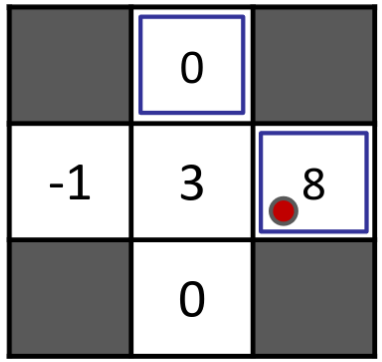
Observed Transitions

B, east, C, -2



C, east, D, -2





$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^{\pi}(s')\right]$$

Q-learning

Q-learning

- TD learning là một phương pháp đánh giá chiến lược di chuyển
- Tuy nhiên phương pháp này không cho phép học chiến lược di chuyển

Q-learning

- TD learning là một phương pháp đánh giá chiến lược di chuyển
- Tuy nhiên phương pháp này không cho phép học chiến lược di chuyển
- Học chiến lược di chuyển

Q-learning

- TD learning là một phương pháp đánh giá chiến lược di chuyển
- Tuy nhiên phương pháp này không cho phép học chiến lược di chuyển
- Học chiến lược di chuyển
- Q-learning

$$\pi(s) = \arg\max Q(s, a)$$

$$Q(s, a) = \sum_{s}^{\prime} T(s, a, s')[R(s, a, s') + \gamma V(s')]$$

Q-learning

$$Q(s,a) = \sum_{s}^{'} T(s,a,s')[R(s,a,s') + \gamma V(s')]$$

- **Chúng ta có thực sự tính được giá trị Q hay không?**

Q-learning

$$Q(s,a) = \sum_{s}^{'} T(s,a,s')[R(s,a,s') + \gamma V(s')]$$

- Lấy mẫu và tính kỳ vọng tương tự như TD learning:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s,a') \right]$$

Khám phá và khai thác

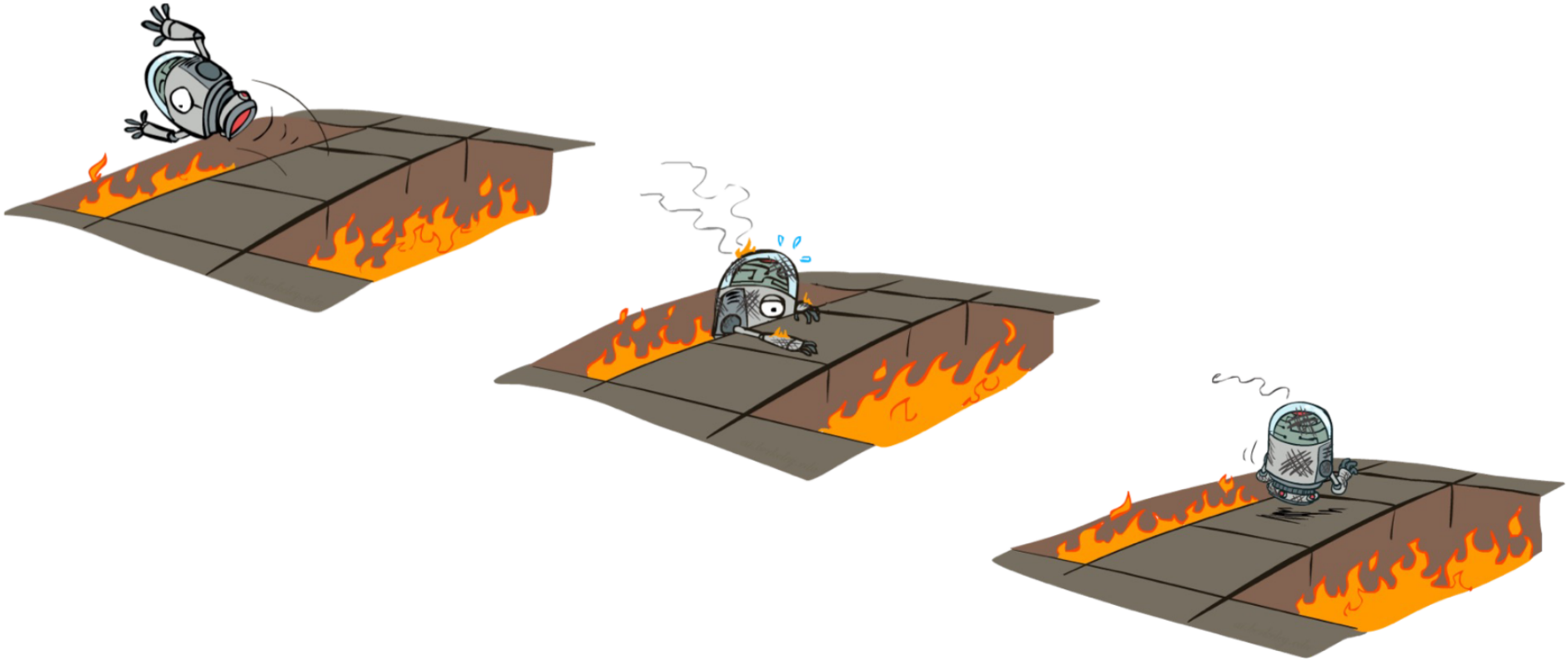
- Một trạng thái có thể có nhiều giá trị đánh giá?

Khám phá và khai thác

- Một trạng thái có thể có nhiều giá trị đánh giá?
- Cần **khám phá** các trạng thái mới
- Gọi giá trị trạng thái là u số lần thăm trạng thái là n sử dụng hàm đánh giá $f(u, n) = u + \frac{k}{n}$

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left[r + \gamma \max_{a'} f\left(Q(s,a'), N(s,a')\right)\right]$$

Active Reinforcement Learning



Ứng dụng

Tài liệu tham khảo

[1] Bùi Tiến Lên, Bộ môn Khoa học máy tính

Bài giảng môn Cơ sở trí tuệ nhân tạo

[2] Michael Negnevitsky

Artificial Intelligence: A Guide to Intelligent Systems (3rd Edition)