The Sarsa Algorithm

$$\begin{array}{cccccc} S & A & R & S & A \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ S_t & A_t & R_{t+1} & S_{t+1} & A_{(t+1)} \end{array}$$

Thuật toán Sarsa đưa ra dự đoán về các giá trị trong cặp hành động trạng thái.

1. Tác nhân chọn hành động ở trạng thái ban đầu để tạo cặp trạng thái-hành động đầu tiên $(S_t \,\&\, A_t)$

2. Thuật toán thực hiện hành động và quan sát phần thưởng tiếp theo và trạng thái tiếp theo $(R_{t+1} \,\&\, S_{t+1})$

3. Tác nhân cam kết hành động tiếp theo của mình trước khi cập nhật ước tính giá trị.

Bằng cách này, thuật toán Sarsa cho phép tác nhân cập nhật ước tính ở mỗi bước thay vì mỗi tập

Phương trình cập nhật Sarsa đầy đủ:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\big)$$

So sánh Sarsa và Q-learning:

- Định nghĩa cơ bản

Tiêu chí	SARSA	Q-learning
Tên đầy đủ	State-Action-Reward-State-**Action**	**Q**-value Learning
Loại thuật toán	**On-policy** (học theo chính sách hiện tại)	**Off-policy** (học theo chính sách tối ưu)

- Tính chất :

+ SARSA : **On-policy**: đánh giá chính sách hiện tại (bao gồm cả các hành động "tệ")

+Q-learning: **Off-policy**: đánh giá chính sách tối ưu (hành động tốt nhất)

- Thái độ với rủi ro:

+ SARSA : Thận trọng hơn (nếu chính sách thăm dò chọn hành động xấu, nó sẽ học điều đó)

+Q-learning: Quyết đoán hơn (luôn học theo hành động tốt nhất)