## Expected Sarsa

- **Ý tưởng chính**:
  Expected Sarsa là một phương pháp TD control tương tự Sarsa và Q-learning, nhưng thay vì lấy mẫu ngẫu nhiên một hành động tiếp theo từ chính sách (như Sarsa), Expected Sarsa **tính toán kỳ vọng giá trị hành động kế tiếp trực tiếp**, bằng cách lấy trung bình trọng số các giá trị hành động theo xác suất chọn hành động của chính sách.

- **Dựa trên Bellman equation cho giá trị hành động**, nhưng Expected Sarsa dùng kỳ vọng (expectation) trên tất cả các hành động kế tiếp, không chỉ lấy mẫu một hành động cụ thể.

- **Công thức cập nhật** giống Sarsa, nhưng phần lỗi TD (temporal difference error) sử dụng giá trị kỳ vọng của hành động tiếp theo thay vì giá trị của hành động được lấy mẫu.

- **Ưu điểm lớn**:

    - Cập nhật ổn định hơn, **độ phương sai thấp hơn** so với Sarsa, vì không phụ thuộc vào mẫu ngẫu nhiên của hành động kế tiếp.

    - Ví dụ minh họa cho thấy trong trường hợp phần thưởng luôn là 1, Sarsa có thể cập nhật sai hướng do mẫu hành động, còn Expected Sarsa luôn cập nhật chính xác.

- **Nhược điểm**:

    - Tính toán kỳ vọng đòi hỏi phải tính trung bình qua tất cả các hành động có thể — nên **tốn thời gian và công sức tính toán hơn**, đặc biệt khi số hành động rất lớn.

**Ba thuật toán chính:**

- **Sarsa:**

    - Dùng phương trình Bellman dạng mẫu (sample-based Bellman equation).

    - Học giá trị hành động $Q_\pi$ (theo chính sách đang thực hiện).

    - Là thuật toán **on-policy** (học theo chính sách hiện tại).

- **Q-learning:**

    - Dùng phương trình Bellman tối ưu (Bellman optimality equation).
    - Học giá trị hành động Q* (giá trị hành động tối ưu).
    - Là thuật toán **off-policy** (học giá trị tối ưu bất kể chính sách đang theo dõi).
- **Expected Sarsa:**
    - Dùng cùng phương trình Bellman như Sarsa nhưng tính mẫu khác.
    - Tính **kỳ vọng** trên các giá trị hành động kế tiếp thay vì lấy mẫu ngẫu nhiên.
    - Có thể dùng cả on-policy và off-policy.

**So sánh On-policy vs Off-policy:**

- **Sarsa** (on-policy) học giá trị cho chính sách đang chạy.
- **Q-learning** (off-policy) học giá trị tối ưu, không phụ thuộc chính sách hiện tại.
- **Expected Sarsa** có thể học theo cả hai cách, linh hoạt hơn.

**Hiệu suất trên bài toán thực tế (ví dụ cliff world):**

- **Sarsa** có thể làm tốt hơn Q-learning trong học trực tuyến vì tính đến việc thăm dò của chính nó.
- Q-learning do thăm dò ngẫu nhiên nên dễ gặp rủi ro (ví dụ: rơi xuống vực).
- Sarsa học được đường đi dài hơn nhưng an toàn hơn.
- **Expected Sarsa** vượt trội hơn Sarsa trong bài toán này vì giảm phương sai do tính kỳ vọng trên hành động kế tiếp.