Tổng quan về Actor-Critic

- **Kết hợp các ý tưởng:** Thuật toán Actor-Critic kết hợp tối ưu hóa chính sách trực tiếp, hàm giá trị và học khác biệt tạm thời (Temporal Difference - TD).
- **Hai thành phần chính:**
    - **Critic (Người phê bình):** Ước tính hàm giá trị (đánh giá chính sách) bằng cách sử dụng TD bán gradient (semi-gradient TD).
    - **Actor (Diễn viên):** Cập nhật các tham số chính sách (dựa trên gradient chính sách) sử dụng TD error (TDR) từ Critic.

Triển khai cụ thể: Softmax Policy với Ưu tiên hành động tuyến tính

Video này tập trung vào việc xây dựng một thuật toán cho các **tập hợp hành động hữu hạn và trạng thái liên tục**.

1. **Tham số hóa chính sách Softmax:**
    - Một lựa chọn phổ biến cho các hành động hữu hạn là chính sách Softmax.

- Chính sách Softmax: $\pi(a|s,\theta) = \frac{\exp(h(s,a,\theta))}{\sum_b \exp(h(s,b,\theta))}$ Đảm bảo xác suất hành động dương và tổng bằng một.

    - Chính sách được viết như một hàm của trạng thái hiện tại, nghĩa là có một phân phối Softmax khác nhau cho mỗi trạng thái.
- **Chọn hành động:** Ở trạng thái hiện tại, chúng ta truy vấn Softmax cho mỗi hành động, tạo ra một vector xác suất. Sau đó, chọn một hành động tỷ lệ thuận với các xác suất này.

Lựa chọn tham số hóa ảnh hưởng đến Hàm giá trị và Ưu tiên hành động:

- **Đặc trưng:** Chúng ta sử dụng **cùng một tập hợp đặc trưng** cho cả ước tính giá trị (Critic) và chính sách (Actor).
- **Critic (Đánh giá giá trị trạng thái):**
    - Cập nhật ước tính hàm giá trị trạng thái (V(s)).
    - Chỉ yêu cầu một **vector đặc trưng đặc trưng cho trạng thái hiện tại (ϕ(s))**.
    - **Ưu tiên hành động của Actor:**
        - Phụ thuộc vào cả trạng thái và hành động.
        - Do đó, hàm ưu tiên hành động đòi hỏi một **vector đặc trưng trạng thái-hành động**.
        - **Giải quyết vấn đề đặc trưng trạng thái-hành động:** Sử dụng chiến lược **xếp chồng đặc trưng trạng thái** (stacked state features). Tức là, một bản sao của vector đặc trưng trạng thái được sử dụng cho mỗi hành động.

- Kích thước của vector tham số chính sách (θ) sẽ lớn hơn trọng số (W) của Critic (ví dụ: nếu có 3 hành động, θ sẽ lớn gấp 3 lần W).

Phương trình cập nhật

1. **Cập nhật của Critic:**
   - **Hàm giá trị tuyến tính:** Gradient của hàm giá trị tuyến tính chỉ là vector đặc trưng.
   - Cập nhật trọng số của Critic: $\mathbf{W} \leftarrow \mathbf{W} + \alpha\delta\phi(S_t)$
     - α: tốc độ học.
     - δ: Sai số TD (TDR) từ Critic.
     - $\phi(S_t)$: Vector đặc trưng trạng thái.
   - Đây chính là TD bán gradient thông thường.
2. **Cập nhật của Actor (Ưu tiên hành động):**
   - Cập nhật tham số chính sách θ của ưu tiên hành động phức tạp hơn một chút.
   - **Gradient của ưu tiên hành động tuyến tính với chính sách Softmax:**
     - Được cung cấp mà không chứng minh: Gradient có hai phần.
     - Phần đầu tiên là các đặc trưng trạng thái-hành động cho hành động được chọn.
     - Phần thứ hai là các đặc trưng trạng thái-hành động được nhân với chính sách (xác suất) tổng hợp trên tất cả các hành động.
   - $\nabla_\theta \ln \pi(A_t|S_t,\theta) = \mathbf{x}(S_t,A_t) - \sum_a \pi(a|S_t,\theta)\mathbf{x}(S_t,a)$
   - Trong đó, $\mathbf{x}(S_t,A_t)$ là vector đặc trưng trạng thái-hành động đã xếp chồng cho hành động A_t được chọn.
- Cập nhật tham số của Actor: $\theta \leftarrow \theta + \alpha\delta\nabla_\theta \ln \pi(A_t|S_t,\theta)$
  - α: tốc độ học.
  - δ: Sai số TD (TDR).
  - $\nabla_\theta \ln \pi(A_t|S_t,\theta)$: Gradient log-xác suất của chính sách đối với tham số.