Xác định và Học Chính sách Trực tiếp (Policy Parameterization)

- **Khái niệm mới:** Thay vì ước tính giá trị hành động (Q-values) rồi chuyển đổi thành chính sách (như GPI), chúng ta sẽ khám phá một lớp phương pháp mới mà trong đó, **chính sách được tham số hóa trực tiếp**.
- **Ví dụ Mountain Car:**
    - Trước đây, chúng ta dùng epsilon-greedy để chuyển đổi giá trị hành động thành chính sách.
    - Tuy nhiên, một chính sách cũng có thể ánh xạ trực tiếp từ trạng thái sang hành động mà không cần tính toán giá trị hành động.
    - Ví dụ đơn giản: Trong Mountain Car, nếu vận tốc dương thì tăng tốc sang phải, ngược lại tăng tốc sang trái. Chính sách này tăng tốc theo hướng đang di chuyển. Đây là một ví dụ về chính sách không dùng giá trị hành động và gần tối ưu.
- **Học chính sách trực tiếp:** Chúng ta sẽ không xác định chính sách thủ công mà sẽ **học chúng**.
    - Sử dụng ngôn ngữ xấp xỉ hàm để biểu diễn và học chính sách trực tiếp.
    - Sử dụng ký hiệu θ (theta) cho vector tham số của chính sách, để phân biệt với W của hàm giá trị.
    - Ký hiệu $\pi(a|s,\theta)$ biểu thị chính sách được tham số hóa: với một trạng thái đầu vào và hành động, nó sẽ xuất ra xác suất thực hiện hành động đó trong trạng thái đó. Ánh xạ này được kiểm soát bởi các tham số θ.

Yêu cầu đối với Chính sách được Tham số hóa

- Chính sách được tham số hóa phải tạo ra một **phân phối xác suất hợp lệ** trên các hành động cho mỗi trạng thái. Điều này có nghĩa là:
    - Xác suất chọn một hành động phải **lớn hơn hoặc bằng 0**.
    - Đối với mỗi trạng thái, **tổng xác suất** của tất cả các hành động phải bằng **một**.
- Điều này đòi hỏi phải suy nghĩ kỹ. Ví dụ, chúng ta không thể dùng hàm tuyến tính trực tiếp như với xấp xỉ hàm giá trị, vì không có cách dễ dàng nào để đảm bảo một hàm tuyến tính sẽ tổng bằng một. Thay vào đó, chúng ta cần hạn chế lớp hàm có thể sử dụng để xây dựng chính sách.

Phân biệt Ưu tiên Hành động và Giá trị Hành động

- **Ưu tiên hành động:** Chỉ ra mức độ tác nhân "ưa thích" mỗi hành động, nhưng chúng **không phải là tóm tắt của phần thưởng trong tương lai**. Chỉ sự khác biệt tương đối giữa các ưu tiên mới quan trọng (ví dụ: thêm 100 vào tất cả các ưu tiên sẽ không thay đổi chính sách).
- **So sánh với Epsilon-greedy:**

    - Chính sách epsilon-greedy (xuất phát từ giá trị hành động) có thể hành xử rất khác so với chính sách softmax trên ưu tiên hành động.
    - Trong epsilon-greedy, hành động có giá trị cao nhất được chọn với xác suất cao, còn các hành động khác có xác suất khá nhỏ. Các hành động có giá trị gần giống nhưng thấp hơn vẫn có xác suất được chọn rất thấp. Tuy nhiên, các hành động có giá trị rất tệ vẫn có thể được chọn thường xuyên do bước thám hiểm epsilon.
    - Ngược lại, chính sách softmax sẽ giảm đáng kể xác suất chọn các hành động có ưu tiên thấp, ngay cả khi nó không bằng không.

Chính sách Softmax

- **Định nghĩa:** Đây là một cách đơn giản nhưng hiệu quả để thỏa mãn các điều kiện của một chính sách hợp lệ.
- **Ưu tiên hành động ($h(s, a, \theta)$):**
    - Là một hàm của trạng thái, hành động và vector tham số θ.
    - Ưu tiên càng cao cho một hành động cụ thể trong một trạng thái có nghĩa là hành động đó có nhiều khả năng được chọn hơn.
    - Ưu tiên hành động có thể được tham số hóa theo bất kỳ cách nào chúng ta muốn (ví dụ: hàm tuyến tính của các đặc trưng trạng thái-hành động, hoặc đầu ra của mạng nơ-ron), vì softmax sẽ đảm bảo các ràng buộc của một phân phối xác suất.
- **Công thức:**

$$\pi(a|s, \theta) = \frac{\exp(h(s, a, \theta))}{\sum_b \exp(h(s, b, \theta))}$$

Hàm mũ (exp) đảm bảo xác suất là dương cho mỗi hành động.

    - Mẫu số (tổng trên tất cả các hành động) chuẩn hóa đầu ra sao cho tổng xác suất của các hành động bằng một.
- **Đặc điểm của Softmax:**
    - Ưu tiên đầu vào có thể lớn tùy ý hoặc thậm chí âm.
    - Nếu một ưu tiên lớn hơn nhiều so với các ưu tiên khác, xác suất hành động sẽ gần bằng một.
    - Nếu một ưu tiên rất nhỏ (ví dụ: âm lớn), chính sách softmax vẫn sẽ chọn hành động đó với xác suất khác không (nhưng rất nhỏ).
    - Các hành động có ưu tiên tương tự sẽ được chọn với xác suất gần bằng nhau.