Mục tiêu cho các thuật toán Policy Gradient

1. **Mục tiêu rõ ràng hơn:**
    - Việc xây dựng một mục tiêu cho việc học một chính sách được tham số hóa trực tiếp theo một nghĩa nào đó thì đơn giản hơn so với các phương pháp dựa trên giá trị hành động.
    - Mục tiêu cuối cùng của học tăng cường là học một chính sách thu được càng nhiều phần thưởng càng tốt về lâu dài. Khi chúng ta tham số hóa chính sách trực tiếp, chúng ta có thể sử dụng trực tiếp mục tiêu này làm mục tiêu học tập.
2. **Các dạng phần thưởng dài hạn:** Trong chuyên môn này, chúng ta đã giới thiệu một vài cách diễn giải khác nhau về ý nghĩa của việc "thu được càng nhiều phần thưởng càng tốt về lâu dài":
    - **Trường hợp theo tập (Episodic case):** Chúng ta sử dụng tổng phần thưởng không chiết khấu (undiscounted return) trong suốt một tập.
    - **Trường hợp liên tục (Continuing case):**
        - **Phần thưởng chiết khấu (Discounted return):** Ưu tiên phần thưởng tức thì để giữ cho tổng hữu hạn.
        - **Phần thưởng trung bình (Average reward formulation):** Tối đa hóa trung bình dài hạn của phần thưởng. Lợi nhuận phù hợp ở đây là tổng của các khác biệt giữa phần thưởng tức thì và phần thưởng trung bình. Vì phần thưởng trung bình dài hạn được trừ đi, tổng này hữu hạn ngay cả khi không có chiết khấu.
3. **Lựa chọn mục tiêu:**
    - Chúng ta có ba công thức vấn đề tiềm năng để xem xét.
    - Chúng ta sẽ giới hạn sự chú ý vào cài đặt liên tục với **phần thưởng trung bình**.
    - Mục tiêu là tìm cách trực tiếp tối ưu hóa các tham số của một chính sách.
4. **Viết mục tiêu phần thưởng trung bình dưới dạng tối ưu hóa:**
    - Trước đây, chúng ta ước tính phần thưởng trung bình để học giá trị hành động trong một biến thể Sarsa có phần thưởng trung bình.
    - Bây giờ, mục tiêu của chúng ta là học một chính sách trực tiếp tối ưu hóa phần thưởng trung bình.
    - Phần thưởng trung bình ($R\pi$) đạt được bởi một chính sách π có thể được viết như sau: $R\pi = s\sum\mu(s)a\sum\pi(a|s)s',r\sum p(s',r|s,a)r$
        - **Tổng trong cùng ($\sum s',r p(s',r|s,a)r$):** Cho phần thưởng kỳ vọng nếu chúng ta bắt đầu ở trạng thái s và thực hiện hành động a.
        - **Tổng ở giữa ($\sum a\pi(a|s)\ldots$):** Cho phần thưởng kỳ vọng dưới chính sách π từ một trạng thái s cụ thể.
        - **Tổng ngoài cùng ($\sum s\mu(s)\ldots$):** Cho phần thưởng trung bình tổng thể bằng cách xem xét tỷ lệ thời gian chúng ta ở trạng thái s dưới chính sách π. Phân phối $\mu(s)$ cung cấp các xác suất này (phân phối trạng thái dừng).

    - Rπ là mục tiêu học tập phần thưởng trung bình của chúng ta.
5. **Phương pháp Policy Gradient (Policy Gradient Methods):**

    - Mục tiêu của tối ưu hóa chính sách sẽ là tìm một chính sách tối đa hóa phần thưởng trung bình.
    - Cách tiếp cận cơ bản của chúng ta sẽ là ước tính **gradient của mục tiêu đối với các tham số chính sách** và điều chỉnh các tham số dựa trên ước tính này.
    - Các lớp phương pháp sử dụng ý tưởng này thường được gọi là **phương pháp policy gradient**.
6. **Sự khác biệt với các phương pháp dựa trên giá trị hành động:**

    - Trước đây, chúng ta sử dụng khung Kế hoạch Chính sách Tổng quát (Generalized Policy Iteration) để học các giá trị hành động xấp xỉ, sau đó gián tiếp suy ra một chính sách tốt từ các giá trị này.
    - Bây giờ, chúng ta quan tâm đến việc học chính sách trực tiếp.
    - Ngoài ra, trước đây chúng ta tối thiểu hóa sai số giá trị bình phương trung bình, còn bây giờ chúng ta tối đa hóa một mục tiêu. Điều đó có nghĩa là chúng ta sẽ muốn di chuyển theo hướng gradient chứ không phải là âm gradient.
7. **Thách thức trong việc tính toán gradient:**

    - Khó khăn chính là việc sửa đổi chính sách của chúng ta sẽ **thay đổi phân phối μ** (phân phối trạng thái dừng).
    - Điều này trái ngược với xấp xỉ hàm giá trị, nơi chúng ta tối thiểu hóa sai số giá trị bình phương trung bình dưới một chính sách cụ thể, và phân phối μ là cố định (nó không thay đổi khi các trọng số trong hàm giá trị được tham số hóa thay đổi). Do đó, chúng ta có thể ước tính gradient cho các trạng thái được rút ra từ μ bằng cách đơn giản tuân theo chính sách.
    - Điều này ít đơn giản hơn vì bản thân μ phụ thuộc vào chính sách mà chúng ta đang tối ưu hóa.
    - May mắn thay, có một câu trả lời lý thuyết tuyệt vời cho thách thức này được gọi là **định lý policy gradient (policy gradient theorem)**, sẽ được thảo luận sau.