Ước tính Gradient Dựa trên Mẫu (Sample-based Gradient Estimate)

Chúng ta muốn dẫn ra một thuật toán giảm gradient cho chính sách của mình. Chúng ta đã có mục tiêu và gradient của nó nhờ định lý policy gradient. Bây giờ, chúng ta cần tìm cách xấp xỉ gradient này. Thực tế, chúng ta sẽ lấy một **mẫu ngẫu nhiên của gradient**.

1. **Từ tổng trên các trạng thái thành kỳ vọng:**
    - Nhắc lại biểu thức gradient của phần thưởng trung bình. Việc tính tổng trên tất cả các trạng thái là không thực tế.
    - Tuy nhiên, chúng ta có thể làm điều tương tự như khi dẫn ra quy tắc giảm gradient ngẫu nhiên cho đánh giá chính sách: chúng ta chỉ cần thực hiện các cập nhật từ các trạng thái mà chúng ta quan sát được khi tuân theo chính sách π.
    - Gradient từ trạng thái St cung cấp một phép xấp xỉ cho gradient của phần thưởng trung bình. Như đã thảo luận trước đây, với giảm gradient ngẫu nhiên, chúng ta có thể điều chỉnh trọng số bằng phép xấp xỉ này và vẫn đảm bảo sẽ đạt đến một điểm dừng.
    - Cập nhật giảm gradient ngẫu nhiên cho các tham số chính sách trông như sau:
        - Các tham số θ được điều chỉnh tỷ lệ với một gradient ngẫu nhiên của mục tiêu.
        - Chúng ta sử dụng tham số kích thước bước α để kiểm soát độ lớn của bước theo hướng đó. α có vai trò tương tự như trong các thuật toán học khác.
2. **Từ tổng trên các hành động thành kỳ vọng:**
    - Hãy xem xét lại biểu thức này từ góc độ kỳ vọng để đơn giản hóa cập nhật và hiểu sâu hơn về lý do tại sao cập nhật này lại có ý nghĩa.
    - Tổng trên các trạng thái được trọng số hóa bởi μ (phân phối dừng) có thể được viết lại dưới dạng một **kỳ vọng dưới μ**. Các trạng thái chúng ta quan sát được khi tuân theo chính sách π được phân phối theo μ.
    - Điều này gợi ý một sự đơn giản hóa khác: bên trong kỳ vọng, chúng ta có một tổng trên tất cả các hành động. Chúng ta muốn làm cho thuật ngữ này đơn giản hơn nữa và loại bỏ tổng trên tất cả các hành động.
    - Nếu đây là một kỳ vọng trên các hành động, chúng ta có thể lấy một mẫu ngẫu nhiên của nó và tránh tổng trên tất cả các hành động.
    - **Mẫu gradient không thiên vị từ một hành động duy nhất:** Để có được ước tính gradient không thiên vị chỉ bằng cách sử dụng **một hành động At** (hành động được tác nhân thực hiện), chúng ta có thể nhân và chia cho $\pi(At|St,\theta)$. Điều này biến tổng có trọng số thành một kỳ vọng trên các hành động được rút ra từ π.
    - Cập nhật tăng gradient ngẫu nhiên mới trông tương tự như công thức đã trình bày, nhưng sử dụng chỉ một hành động At.

3. **Mẹo toán học: Gradient của logarit tự nhiên:**
    - Thông thường, gradient được viết lại dưới dạng **gradient của logarit tự nhiên của** π ($\nabla_\theta \ln \pi(A_t|S_t,\theta)$).
    - Điều này dựa trên một công thức từ giải tích: $\nabla_\theta \pi(a|s,\theta) = \pi(a|s,\theta)\nabla_\theta \ln \pi(a|s,\theta)$.
    - **Lý do:**
        - Đôi khi việc tính gradient của logarit của một số phân phối thực sự đơn giản hơn.
        - Nó cho phép viết gradient một cách gọn hơn.
    - Đây chỉ là một thủ thuật toán học và không làm thay đổi thuật toán cơ bản.

Tính toán Gradient ngẫu nhiên

Chúng ta hiện đã có một quy tắc cập nhật rõ ràng để học các tham số chính sách. Để thực sự tính toán gradient ngẫu nhiên cho một trạng thái và hành động nhất định, chúng ta chỉ cần hai thành phần:

1. **Gradient của chính sách ($\nabla_\theta \pi(A_t|S_t,\theta)$):** Chúng ta biết chính sách và cách tham số hóa của nó, vì vậy có thể tính toán gradient này một cách trực tiếp.
2. **Ước tính giá trị vi phân (differential values):** Giá trị hành động có thể được xấp xỉ bằng nhiều cách. Ví dụ, chúng ta có thể sử dụng một thuật toán TD học các giá trị hành động vi phân.

1.