Thiết lập Phần thưởng Trung bình

- **Vấn đề với Chiết khấu:** Trong các tác vụ liên tục, việc chiết khấu (sử dụng $\gamma<1$) giúp cân bằng giữa lợi ích ngắn hạn và dài hạn. Tuy nhiên, việc lựa chọn γ phụ thuộc vào bài toán; một giá trị γ phù hợp với bài toán này (ví dụ: vòng ngắn) có thể quá nhỏ đối với bài toán khác (ví dụ: vòng dài), có khả năng dẫn đến các lựa chọn chính sách dưới tối ưu về lâu dài. Để đảm bảo các hành động tối đa hóa phần thưởng dài hạn, γ cần phải tiến gần đến 1, điều này có thể dẫn đến phần thưởng lớn, biến thiên và có khả năng vô hạn, gây khó khăn cho việc học.
- **Công thức Phần thưởng Trung bình ($\bar{R}^{\pi}$):** Cách tiếp cận này tập trung vào việc tối đa hóa **phần thưởng trung bình mỗi bước thời gian** trong một chân trời dài. Nó coi trọng phần thưởng ngay lập tức và phần thưởng xa ngang nhau, thể hiện *tỷ lệ phần thưởng* mà tác nhân nhận được.
    - Về mặt toán học, đó là kỳ vọng của phần thưởng trên các trạng thái, được trọng số hóa theo tần suất mỗi trạng thái được ghé thăm dưới chính sách.
- **Ví dụ MDP Cận thị (Nearsighted MDP):**
    - Một MDP với hai vòng giao nhau và một điểm quyết định (trạng thái S).
    - **Vấn đề Chiết khấu:** Đối với $\gamma=0.5$, chính sách chọn vòng bên trái (phần thưởng +1 sau 5 bước) được ưu tiên. Đối với $\gamma=0.9$, chính sách chọn vòng bên phải (phần thưởng +2 sau 5 bước) được ưu tiên. Chính sách được ưu tiên phụ thuộc rất nhiều vào γ. Đối với các vòng dài hơn, γ sẽ cần phải rất gần 1.
    - **Giải pháp Phần thưởng Trung bình:**
        - Vòng bên trái: Phần thưởng trung bình là $1/5=0.2$.
        - Vòng bên phải: Phần thưởng trung bình là $2/5=0.4$.
        - Phần thưởng trung bình rõ ràng ưu tiên vòng bên phải ($0.4>0.2$) bất kể khái niệm "chiết khấu", phản ánh trực tiếp tổng phần thưởng thu được theo thời gian.

Hàm Giá trị Vi phân (Differential Value Functions)

- **Nhu cầu về Giá trị Hành động:** Mặc dù phần thưởng trung bình cho chúng ta biết chính sách nào tốt hơn, chúng ta cần một cách để xác định hành động nào tốt hơn từ một trạng thái nhất định trong khuôn khổ này. Điều này dẫn đến **lợi nhuận vi phân (differential returns)**.
- **Lợi nhuận Vi phân:** Được định nghĩa là tổng của **(phần thưởng - phần thưởng trung bình $\bar{R}^{\pi}$)** cho mỗi bước thời gian. Nó thể hiện *lượng phần thưởng bổ sung* mà tác nhân sẽ nhận được từ trạng thái-hành động hiện tại so với phần thưởng trung bình của chính sách.
    1. Đây là một tổng hội tụ chỉ khi hằng số trừ đi chính xác bằng phần thưởng trung bình thực.
- **Hàm Giá trị Vi phân ($V^{\pi}(s)$, $Q^{\pi}(s,a)$):** Được định nghĩa là lợi nhuận vi phân kỳ vọng.

1. $V_\pi(s)$ nắm bắt lượng phần thưởng bổ sung mà tác nhân nhận được bằng cách bắt đầu ở trạng thái s so với phần thưởng trung bình của nó theo chính sách π.
2. $Q_\pi(s,a)$ nắm bắt điều tương tự khi bắt đầu bằng hành động a ở trạng thái s.

- **Phương trình Bellman cho Hàm Giá trị Vi phân:** Chúng trông tương tự như các phương trình Bellman được chiết khấu nhưng có hai điểm khác biệt chính:
    1. Phần thưởng tức thì (R_{t+1}) bị trừ đi phần thưởng trung bình $(\bar{R}_\pi)$.
    2. **Không có chiết khấu** $(\gamma=1)$.

Các Thuật toán cho Phần thưởng Trung bình

- Nhiều thuật toán từ cài đặt chiết khấu có thể được điều chỉnh.
- **Sarsa Vi phân (Differential Sarsa):**
    - Rất giống với thuật toán Sarsa tiêu chuẩn.
    - **Điểm khác biệt chính:** Nó phải **theo dõi ước tính phần thưởng trung bình $(\bar{R})$ dưới chính sách hiện tại của nó** và trừ ước tính này khỏi phần thưởng mẫu trong quy tắc cập nhật của nó. Điều này thường được thực hiện bằng cách sử dụng các kỹ thuật trung bình gia tăng.
    - Trong thực tế, một bản cập nhật sửa đổi cho $\bar{R}$ với phương sai thấp hơn (so với trung bình theo cấp số nhân) thường được sử dụng để có hiệu suất tốt hơn.