



2. Giá trị của một hành động (Action Value)

- **Định nghĩa:** Giá trị hành động $q(a)$ là "kỳ vọng" của reward mà tác nhân sẽ nhận được khi thực hiện hành động a . Nói cách khác:

$$q_*(a) = E[R \mid A = a]$$

trong đó R là reward thu được ngay sau khi chọn hành động $A = a$. 

- **Thực tế:** Ta không biết trước phân phối reward cho mỗi hành động (khi bắt đầu). Do đó, $q(a)$ là một giá trị ẩn cần ước lượng từ dữ liệu quan sát. 

3. Phương pháp mẫu trung bình (Sample-Average Method)

- **Ý tưởng cơ bản:**
 1. Mỗi khi chọn hành động a , ta quan sát được reward R_t .
 2. Ta cộng dồn tổng reward đã nhận được khi chọn a .
 3. Khi cần ước lượng $q(a)$, ta chia tổng reward của a cho số lần a đã được chọn.

- **Công thức:**
Giả sử sau n lần lựa a , tổng reward quan sát được là


$$S_n(a) = \sum_{i=1}^n R_i(a).$$

Khi đó, ước lượng mẫu trung bình là



$$\hat{q}_n(a) = \frac{S_n(a)}{n}.$$

Mỗi khi có thêm reward $R_{n+1}(a)$, ta cập nhật:


$$S_{n+1}(a) = S_n(a) + R_{n+1}(a), \quad \hat{q}_{n+1}(a) = \frac{S_{n+1}(a)}{n+1}.$$

- **Ưu điểm:** Đơn giản, hội tụ về giá trị thật của $q_*(a)$ nếu reward độc lập và phân phối cố định.
- **Nhược điểm:**
 - Cần lưu lại tổng reward và số lần chọn mỗi hành động.
 - Nếu môi trường thay đổi (nonstationary), phương pháp cộng dồn này không kịp thích ứng với sự thay đổi do trọng số đã gom từ quá khứ. 


4. Ví dụ minh họa: Thử nghiệm lâm sàng (Clinical Trials)


- **Bối cảnh:** Bác sĩ phải quyết định kê một trong ba phương pháp điều trị (Treatment P, Q, R) cho mỗi bệnh nhân.
 - Nếu bệnh nhân khỏi bệnh, bác sĩ chấm reward = 1.
 - Nếu không cải thiện, reward = 0. 
- **Khởi tạo:** Ban đầu, bác sĩ chưa có thông tin gì, nên ước lượng giá trị cho ba phương pháp đều bằng 0 (tức $\hat{q}_0(P) = \hat{q}_0(Q) = \hat{q}_0(R) = 0$). 
- **Bệnh nhân thứ nhất:**
 1. Bác sĩ chọn ngẫu nhiên (hoặc theo chính sách ban đầu) phương pháp P.
 2. Bệnh nhân khỏi, thu được reward $R = 1$.
 3. Cập nhật: tổng reward cho P từ 0 \rightarrow 1, số lần chọn P từ 0 \rightarrow 1, nên

$$\hat{q}_1(P) = \frac{1}{1} = 1.$$
- **Bệnh nhân thứ hai:**
 1. Bác sĩ (theo chiến lược có thể là ϵ -greedy hoặc vẫn chọn ngẫu nhiên) tiếp tục kê P.
 2. Lần này P không hiệu quả, reward = 0.
 3. Cập nhật: tổng reward cho P từ 1 \rightarrow 1, số lần chọn P từ 1 \rightarrow 2, nên



$$\hat{q}_2(P) = \frac{1}{2} = 0.5.$$
- **Sau nhiều bệnh nhân:**
 - Khi mỗi phương pháp được thử đủ, bác sĩ có thể tính ước lượng $\hat{q}(P)$, $\hat{q}(Q)$, $\hat{q}(R)$ dựa trên dữ liệu thực nghiệm.
 - Theo thời gian, ước lượng hội tụ gần đến giá trị thật $q_*(a)$, giúp xác định phương pháp tốt nhất. 

5. Phương pháp Khai thác (Greedy Method)





- **Định nghĩa:**
 - “Greedy action” là hành động (cần gặt, hay phương pháp điều trị) có ước lượng giá trị cao nhất tại thời điểm hiện tại.
 - Khi chọn greedy action, ta đang **khai thác** (exploitation) kiến thức hiện có, mong muốn thu được reward ngay lập tức dựa trên ước lượng tốt nhất hiện tại. 
- **Hạn chế:**
 - Luôn chọn hành động có $\hat{q}(a)$ lớn nhất khiến ta **không thu thập thông tin** về các hành động khác – có thể bỏ qua hành động tốt hơn mà ước lượng hiện tại còn thấp do thiếu dữ liệu.

- Kết quả: dễ rơi vào “cực tiểu địa phương” (local optimum) nếu khởi đầu may mắn chọn trúng hành động không tốt. 
-

6. Khám phá (Exploration)

- **Nhu cầu khám phá:**
 - Để tránh bỏ sót hành động có giá trị thật cao, tác nhân thỉnh thoảng phải chọn những hành động chưa được “tin tưởng” nhiều (ví dụ hành động có $Q(a)$ thấp hoặc mới ít thử).
 - Mục đích: **thu thập thêm thông tin** về các hành động khác, có thể dẫn đến phát hiện hành động tốt hơn trong tương lai. 
 - **Ví dụ chính sách ϵ -greedy:**
 - Với xác suất $1 - \epsilon$, chọn greedy action (khai thác).
 - Với xác suất ϵ , chọn ngẫu nhiên một hành động bất kỳ (khám phá).
 - ϵ thường là số nhỏ (ví dụ 0.1), giúp cân bằng giữa khai thác-khám phá. 
-

7. Tổng kết nhanh

1. **Action value ($Q(a)$):** Kỳ vọng reward khi thực hiện hành động a . Phải ước lượng từ dữ liệu vì phân phối reward chưa biết trước. 
 2. **Sample-average method:** Tính $\hat{Q}_n(a) = \frac{1}{n} \sum_{i=1}^n R_i(a)$. Mỗi lần thu được reward mới, cập nhật trung bình mẫu. 
 3. **Ví dụ lâm sàng:** Bác sĩ thử nhiều lần ba phương pháp, ghi nhận reward 1/0, sau cùng ước lượng giá trị từng phương pháp để chọn ra phương án tốt nhất. 
 4. **Greedy vs. Exploration:**
 - Greedy action: luôn chọn hành động có $Q(a)$ lớn nhất (khai thác).
 - Khám phá: thỉnh thoảng chọn ngẫu nhiên để thu thêm thông tin, tránh bỏ sót hành động tiềm năng. 
-

Gợi ý mở rộng:

- Ngoài sample-average, còn có cách cập nhật giá trị theo kiểu “trung bình lũy tiến” (incremental update) để không cần lưu toàn bộ tổng reward/counter:

$$\hat{Q}_{n+1}(a) = \hat{Q}_n(a) + \frac{1}{n+1}(R_{n+1}(a) - \hat{Q}_n(a)).$$

- Khi môi trường không ổn định (nonstationary), người ta thường dùng ứng suất giảm dần (constant step-size) thay vì $\frac{1}{n+1}$, cho phép nhanh chóng thích ứng với sự thay đổi.
- Các chính sách tiên tiến hơn (như UCB, Thompson Sampling) cân bằng khai thác-khám phá theo những cách khác nhau, thường đạt hiệu quả cao hơn ϵ -greedy.