

1. Định nghĩa Reward

- **Reward (phần thưởng)** là tín hiệu phản hồi dạng vô hướng (scalar) do môi trường cung cấp cho tác nhân (agent) sau khi tác nhân thực hiện một hành động tại một trạng thái và bước thời gian cụ thể.
 - Reward phản ánh mức độ "tốt" của hành động vừa thực hiện, là cơ chế chính để hướng dẫn tác nhân học cách tối ưu hóa mục tiêu. 📄
-

2. Bản chất thời gian của K-Armed Bandit

- Trong nhiều bài toán tăng cường (reinforcement learning), tác nhân tương tác với môi trường qua nhiều bước, trạng thái thay đổi và hành động của tác nhân ảnh hưởng đến trạng thái tiếp theo. Tuy nhiên, **trong bài toán K-Armed Bandit, không có khái niệm trạng thái kế tiếp sau khi chọn hành động.**
 - Cụ thể:
 1. Tác nhân quan sát một "lần duy nhất" (initial observation),
 2. Chọn một hành động (một trong K cần gạt),
 3. Nhận ngay reward từ môi trường,
 4. Vòng đời (episode) kết thúc.
 - Môi trường **không có động lực nội tại** (no dynamics): reward chỉ phụ thuộc vào hành động hiện tại, không phụ thuộc vào quá khứ; không có quá trình chuyển trạng thái liên tục giữa các bước. 📄
-

3. Bài toán K-Armed Bandit

- **K-Armed Bandit** là một ví dụ cổ điển trong lý thuyết quyết định và RL (reinforcement learning).
- Mô tả tình huống:
 - Giả sử có K cần gạt (slot machines) đặt cạnh nhau. Mỗi cần gạt k sở hữu một phân phối xác suất ẩn $P_k(r)$ mô tả khả năng cho reward r .
 - Một "con bạc" (agent) có T lượt kéo (pulls). Mỗi lượt, con bạc chọn một trong K cần gạt và nhận một reward (một số điểm) từ cần gạt đó. Mục tiêu là **tối đa hóa tổng reward** thu được sau T lượt. 📄
- Thách thức chính:
 - **Khám phá (Exploration)**: thử các cần gạt ít rõ ràng để thu thập thông tin về phân phối reward của chúng.
 - **Khai thác (Exploitation)**: tập trung kéo những cần gạt đã biết là có reward kỳ vọng cao từ trước.

- Cân bằng giữa hai yếu tố này (exploration–exploitation) là cốt lõi để đạt tổng reward cao nhất. 📄

4. Định nghĩa Action-Value (Giá trị hành động)

- Để quyết định cần gạt nào là “tốt nhất”, ta gán cho mỗi hành động a (tức kéo cần gạt a) một giá trị kỳ vọng gọi là $q_*(a)$.
- **Action-value function** $q_*(a)$ được định nghĩa là:

$$q_*(a) = E[R_t | A_t = a]$$

trong đó:

- A_t là hành động được chọn ở lượt t (tại bước thời gian t),
- R_t là reward nhận được sau khi thực hiện A_t .
- Thực chất, $q_*(a)$ là “kỳ vọng” của reward khi chọn a , tức:

$$q_*(a) = \sum_r r \cdot P(R_t = r | A_t = a).$$

Nói cách khác, nhân mỗi giá trị reward có thể nhận được với xác suất nó xảy ra (theo phân phối ẩn của cần gạt), rồi cộng lại. 📄





5. Ví dụ minh họa tính Action-Value

Lưu ý: Slide chỉ ghi “Example for calculating action value” nhưng không đưa chi tiết cụ thể. Dưới đây là ví dụ minh họa thêm để làm rõ ý niệm.

- Giả sử có 3 cần gạt ($K=3$), mỗi cần gạt a cho reward là 0 hoặc 1. Định nghĩa:
 - Cần gạt 1: $P(R = 1 | A = 1) = 0.2$, $P(R = 0 | A = 1) = 0.8$.
 - Cần gạt 2: $P(R = 1 | A = 2) = 0.5$, $P(R = 0 | A = 2) = 0.5$.
 - Cần gạt 3: $P(R = 1 | A = 3) = 0.8$, $P(R = 0 | A = 3) = 0.2$.
- **Tính giá trị $q_*(a)$ cho mỗi cần gạt:**
 1. $q_*(1) = 0 \times 0.8 + 1 \times 0.2 = 0.2$.
 2. $q_*(2) = 0 \times 0.5 + 1 \times 0.5 = 0.5$.
 3. $q_*(3) = 0 \times 0.2 + 1 \times 0.8 = 0.8$.
- **Kết luận:** Cần gạt số 3 đang có **giá trị kỳ vọng cao nhất** (0.8), vì vậy nếu chỉ khai thác (exploitation), tác nhân sẽ luôn kéo cần gạt 3. Tuy nhiên, nếu chưa chắc chắn phân phối của

cần gạt nào, cần cân nhắc khám phá thêm.

6. Tổng kết (Summary của slide)

- **Define reward:** Reward là tín hiệu vô hướng cho biết mức độ thành công của hành động. 
 - **Understand the temporal nature of the bandit problem:** Bài toán bandit không có "động lực" (dynamics); mỗi lần chọn là độc lập, không ảnh hưởng trạng thái kế tiếp. 
 - **Define K-Armed Bandit:** Bài toán con bạc với K cần gạt, mỗi cần gạt có phân phối reward ẩn. Mục tiêu tối đa tổng reward qua một số lượt kéo. 
 - **Define action-values:** Giá trị hành động $q_*(a)$ là kỳ vọng reward khi chọn hành động a , được tính bằng tổng trọng số của mọi giá trị reward có thể với xác suất xảy ra tương ứng. 
-

Gợi ý áp dụng thêm:

- Thực tế, khi không biết trước $q_*(a)$, ta phải ước lượng giá trị này dần dần qua quá trình "kéo và cập nhật" (ví dụ dùng trung bình mẫu hoặc thuật toán ϵ -greedy, UCB, Thompson Sampling...).
- Bài toán K-Armed Bandit là nền tảng để hiểu sâu hơn về các thuật toán tối ưu hoá khai thác-khám phá, từ đó mở rộng sang các bài toán RL phức tạp hơn (có trạng thái, chuỗi thời gian, Markov Decision Process).