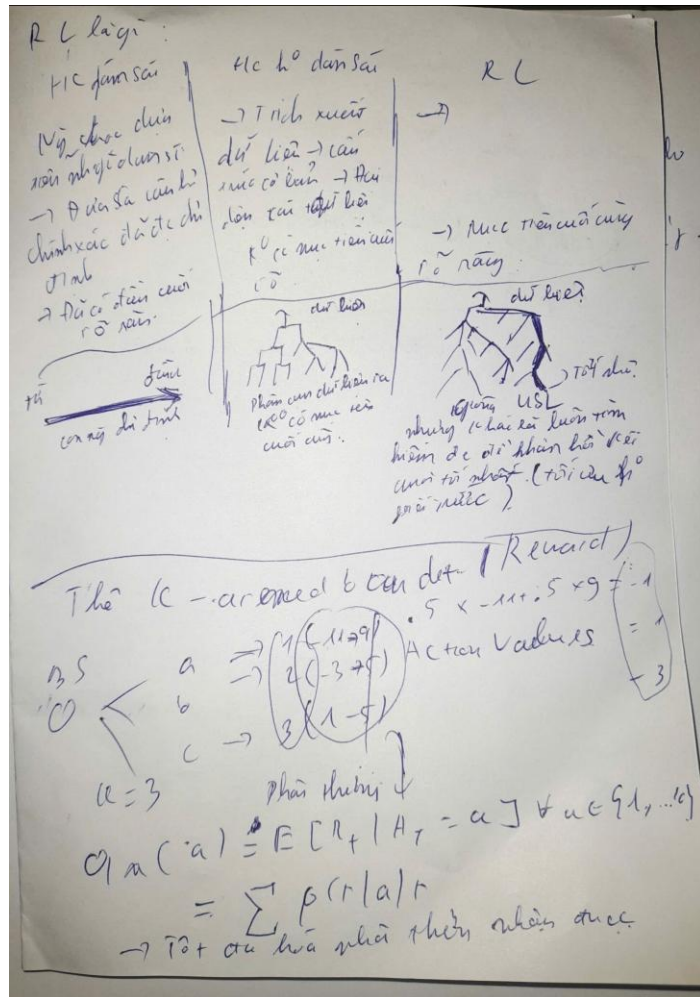


## Fundamentals of Reinforcement Learning

Giới thiệu (Đoạn này em chép tay)



## Module 1 Learning Objectives

By the end of this module, you should be able to meet the following learning objectives:

### Lesson 1: The K-Armed Bandit Problem

- Define reward
- Understand the temporal nature of the bandit problem
- Define k-armed bandit
- Define action-values

### Lesson 2: What to Learn? Estimating Action Values

- Define action-value estimation methods
- Define exploration and exploitation
- Select actions greedily using an action-value function
- Define online learning
- Understand a simple online sample-average action-value estimation method
- Define the general online update equation
- Understand why we might use a constant step-size in the case of non-stationarity

### Lesson 3: Exploration vs. Exploitation Tradeoff

- Define epsilon-greedy
- Compare the short-term benefits of exploitation and the long-term benefits of exploration
- Understand optimistic initial values
- Describe the benefits of optimistic initial values for early exploration
- Explain the criticisms of optimistic initial values
- Describe the upper confidence bound action selection method
- Define optimism in the face of uncertainty

# The K-Armed Bandit Problem

## (1) Define reward: - phần thưởng

**Định nghĩa:** Giá trị (số thực) nhận được ngay sau khi thực hiện một hành động. Phần thưởng phản ánh mức độ "tốt" hay "hiệu quả" của hành động đó.

Công thức:

Ký hiệu phần thưởng tại thời điểm  $t$  là  $R_t$ .

Không có công thức cụ thể, nhưng thường giả định:

$$R_t = \mathbb{E}[r|A_t = a]$$

Nghĩa là kỳ vọng phần thưởng khi chọn hành động  $a$  tại thời điểm  $t$ .

- Ví dụ:
  - ❓ Bác sĩ có 3 loại thuốc (A, B, C) điều trị một bệnh.
- Khi chọn thuốc A cho bệnh nhân, bệnh nhân hồi phục nhanh  $\rightarrow$  phần thưởng  $R_t = +10$ .
- Nếu bệnh nhân gặp tác dụng phụ nhẹ  $\rightarrow R_t = -2$ .

## (2) Temporal property - tính chất thời gian của vấn đề bandit

Định nghĩa: Hành động và phần thưởng xảy ra tuần tự theo thời gian. Lựa chọn hiện tại ảnh hưởng đến kết quả và chiến lược trong tương lai.

Công thức:

Không phải công thức cụ thể, mà là khái niệm sequence of actions and rewards

$$A_1, R_1, A_2, R_2, \dots, A_t, R_t$$

Trong đó:

- $A_t$  là hành động tại thời điểm  $t$
- $R_t$  là phần thưởng nhận được ngay sau  $A_t$

**Ví dụ :**

- Ngày 1 bác sĩ kê thuốc A  $\rightarrow$  bệnh nhân cải thiện nhẹ ( $R_1 = +5$ )
- Ngày 2 dựa vào kết quả hôm trước, bác sĩ đổi sang thuốc B  $\rightarrow$  bệnh nhân hồi phục tốt hơn ( $R_2 = +8$ )
- Quyết định hôm nay ảnh hưởng đến lựa chọn ngày mai.

## (3) K-Armed Bandit

**Định nghĩa:** Mô hình chọn lựa giữa  $K$  hành động (arms), mỗi hành động cho phần thưởng ngẫu nhiên với phân phối khác nhau.

Công thức:

Với mỗi hành động  $a \in \{1, 2, \dots, K\}$ , có phần thưởng ngẫu nhiên theo phân phối  $P_a(r)$

– Mục tiêu: Tối đa hóa tổng phần thưởng kỳ vọng

$$\max \sum_{t=1}^T R_t$$

**Ví dụ:**

- Bác sĩ có 4 phương pháp điều trị ( $K = 4$ ): **thuốc A, thuốc B, phẫu thuật, vật lý trị liệu.**

- Mỗi phương pháp cho phần thưởng khác nhau tùy bệnh nhân (ví dụ thuốc A: hồi phục tốt nhưng lâu; phẫu thuật: hồi phục nhanh nhưng rủi ro cao)
- Bác sĩ phải thử nghiệm và chọn phương pháp tốt nhất cho nhóm bệnh nhân hiện tại.

#### (4) Action Value - Giá trị hành động

**Định nghĩa:** Kỳ vọng phần thưởng nhận được nếu luôn chọn hành động đó.

Công thức (Giá trị thực của hành động  $a$ ):

$$q_*(a) = \mathbb{E}[R_t | A_t = a]$$

Ước lượng giá trị hành động (sample average estimate)

Nếu đã chọn hành động  $a$   $N(a)$  lần và nhận phần thưởng  $r_1, r_2, \dots, r_n$

$$Q_n(a) = \frac{r_1 + r_2 + \dots + r_n}{N(a)}$$

- Bác sĩ đã kê thuốc A cho 5 bệnh nhân — phần thưởng: +6, +7, +5, +8, +6

$$Q(A) = \frac{6 + 7 + 5 + 8 + 6}{5} = 6.4$$

— Giá trị hành động của thuốc A là 6.4, nghĩa là kỳ vọng hồi phục "trung bình tốt" nếu tiếp tục chọn thuốc A.

#### Tóm tắt dễ nhớ

Khái niệm	Công thức	Ví dụ bác sĩ
Phần thưởng (Reward)	$R_t$	Hiệu quả điều trị sau 1 lần kê thuốc
Tính chất thời gian	$A_1, R_1, A_2, R_2, \dots$	Kê thuốc hôm nay ảnh hưởng điều trị ngày mai
K-Armed Bandit	$\max \sum R_t$	Chọn tốt nhất giữa nhiều phương pháp điều trị
Giá trị hành động	$q_*(a), Q_n(a)$	Hiệu quả trung bình của một loại thuốc/phương pháp

## Estimating Action Values

### (1) Define action-value estimation methods

**Định nghĩa:** Cách tính toán giá trị (kỳ vọng phần thưởng) của từng hành động dựa trên phần thưởng thu thập được từ kinh nghiệm.

Công thức (trung bình phần thưởng):

$$Q_n(a) = \frac{r_1 + r_2 + \dots + r_n}{N(a)}$$

Trong đó:

- $Q_n(a)$  là ước lượng giá trị hành động  $a$  sau  $n$  lần chọn
- $r_1, r_2, \dots, r_n$  là phần thưởng nhận được khi chọn hành động  $a$
- $N(a)$  là số lần đã chọn hành động  $a$

**Ví dụ:**

- Kê thuốc B cho 3 bệnh nhân → phần thưởng: +7, +6, +8

$$Q(B) = \frac{7 + 6 + 8}{3} = 7$$

→ Giá trị thuốc B là 7

## (2) Define exploration and exploitation

**Định nghĩa:**

- **Khai thác (Exploitation):** Chọn hành động có giá trị cao nhất hiện tại
- **Khám phá (Exploration):** Thử hành động mới để thu thập thêm thông tin

**Ví dụ:**

**Ví dụ bác sĩ:**

- Thuốc A có  $Q = 6.4$ , thuốc B có  $Q = 7$
- Bác sĩ thường chọn thuốc B (khai thác), nhưng đôi khi thử thuốc C mới (khám phá) để biết thêm hiệu quả

## (3) Select actions greedily using an action-value function

**Định nghĩa:** Luôn chọn hành động có giá trị ước lượng cao nhất

$$A_t = \arg \max_a Q(a)$$

**Ví dụ:**

- Giữa thuốc A ( $Q = 6.4$ ), thuốc B ( $Q = 7$ ), thuốc C ( $Q = 5$ )  
→ Bác sĩ **luôn chọn thuốc B**

## (4) Define online learning

**Định nghĩa:** Cập nhật giá trị hành động liên tục khi có thêm dữ liệu, thay vì đợi đến cuối

**Ví dụ:**

## (5) Understand a simple online sample-average action-value estimation method

**Định nghĩa:** Cập nhật giá trị hành động bằng trung bình cộng các phần thưởng đã thu thập

Công thức cập nhật (đệ quy):

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

- $R_n$  là phần thưởng mới nhất
- $n$  là số lần chọn hành động

**Ví dụ:**

- Đã kê thuốc A 3 lần,  $Q_3 = 6.4$
- Lần 4 phần thưởng  $R_4 = 9$

$$Q_4 = 6.4 + \frac{1}{4}(9 - 6.4) = 6.4 + 0.65 = 7.05$$

→ Giá trị thuốc A cập nhật thành 7.05

## Tóm tắt dễ nhớ

Khái niệm	Công thức	Ý nghĩa bác sĩ
Ước lượng giá trị	$Q_n(a) = \frac{\sum r}{N(a)}$	Trung bình hiệu quả thuốc
Khám phá & khai thác	$\epsilon$ -greedy	Đôi khi thử thuốc mới
Hành động tham lam	$A = \arg \max Q(a)$	Luôn chọn thuốc tốt nhất hiện tại
Học trực tuyến	Cập nhật liên tục	Điều chỉnh liên tục khi điều trị
Trung bình mẫu	$Q_{n+1} = Q_n + \frac{1}{n}(R - Q_n)$	Cập nhật giá trị từ phần thưởng mới
Cập nhật tổng quát	$Q_{n+1} = Q_n + \alpha(R - Q_n)$	Cập nhật mềm dẻo với hệ số $\alpha$
Bước cập nhật không đổi	$\alpha$ cố định	Theo kịp khi bệnh thay đổi

## Lesson 3: Exploration vs. Exploitation Tradeoff

### 1 Epsilon-Greedy ( $\epsilon$ -Greedy)

**Định nghĩa:**

Hầu hết thời gian chọn hành động tốt nhất (khai thác), nhưng thỉnh thoảng ( $\epsilon\%$ ) chọn ngẫu nhiên (khám phá)

Công thức chọn hành động:

$$A_t = \begin{cases} \arg \max_a Q(a) & \text{với xác suất } 1 - \epsilon \\ \text{random action} & \text{với xác suất } \epsilon \end{cases}$$

Ví dụ bác sĩ:

- $\epsilon = 0.1$  (10%)
  - 90% thời gian chọn thuốc tốt nhất
  - 10% thử ngẫu nhiên thuốc khác

### 2 Lợi ích ngắn hạn của khai thác và lợi ích dài hạn của khám phá

**Khai thác (Exploitation):**

Tối đa hóa phần thưởng ngắn hạn bằng cách chọn hành động tốt nhất hiện tại

**Khám phá (Exploration):**

Tìm hiểu thêm → phát hiện hành động tốt hơn → tối ưu dài hạn

Ví dụ bác sĩ:

- Bác sĩ đang dùng thuốc B ( $Q = 7$ ) tốt nhất hiện tại
- Nhưng nếu thử thuốc C vài lần, có thể phát hiện thuốc C có  $Q = 9$  → tốt hơn về lâu dài

### 3 Giá trị khởi tạo lạc quan (Optimistic Initial Values)

Định nghĩa: Gán giá trị ước lượng cao bất thường cho tất cả hành động lúc đầu → thúc đẩy khám phá

Ví dụ bác sĩ:

- 3 loại thuốc (A, B, C) → ban đầu gán  $Q(A) = Q(B) = Q(C) = 10$
- Khi kê thuốc lần đầu chưa ai biết chính xác hiệu quả, nhưng giá trị cao khiến bác sĩ muốn thử hết

### 4 Lợi ích của giá trị khởi tạo lạc quan

Giúp đảm bảo mọi hành động được thử ít nhất 1 lần, tránh bỏ sót lựa chọn tốt

Ví dụ bác sĩ:

- Nếu ban đầu  $Q(A) = Q(B) = 0$ ,  $Q(C) = 0$  → bác sĩ có thể thiên về 1 thuốc đã thử sớm
- Giá trị khởi tạo cao khiến bác sĩ "có động lực" thử đều cả 3 thuốc

### 5 Phê bình giá trị khởi tạo lạc quan

Sau khi khám phá đủ, giá trị ước lượng hội tụ → lợi ích của giá trị khởi tạo lạc quan mất dần  
→ Không còn thúc đẩy khám phá nữa

Ví dụ bác sĩ:

- Sau 10 bệnh nhân, bác sĩ đã thử hết 3 thuốc → biết chắc thuốc B là tốt nhất  
→ Lúc này khởi tạo lạc quan không còn tác dụng

### 6 Phương pháp chọn hành động dựa trên giới hạn tin cậy trên (Upper Confidence Bound - UCB)

Định nghĩa:

Chọn hành động có giá trị lớn nhất xét cả giá trị trung bình và mức độ chưa chắc chắn

Công thức:

$$A_t = \arg \max_a \left( Q(a) + c \sqrt{\frac{\ln t}{N(a)}} \right)$$

- $Q(a)$ : Giá trị trung bình ước lượng
- $N(a)$ : Số lần chọn a
- $t$ : Tổng số lượt chọn
- $c$ : Hệ số điều chỉnh mức độ khám phá

→ Khi  $N(a)$  nhỏ (ít thử) → phần bù lớn → khuyến khích thử hành động đó

Ví dụ bác sĩ:

- Thuốc A (đã kê nhiều lần) → phần bù nhỏ
- Thuốc C (mới thử 1 lần) → phần bù lớn → bác sĩ sẽ thử lại thêm lần nữa

### 7 Lạc quan trong sự không chắc chắn (Optimism in the face of uncertainty)

Định nghĩa: Đánh giá tích cực các hành động chưa thử nhiều để tăng động lực khám phá

→ Đây là tư tưởng chung đằng sau:

- Giá trị khởi tạo lạc quan
- UCB

Ví dụ bác sĩ:

- Bác sĩ tin rằng "biết đâu thuốc ít thử lại rất tốt", nên ưu tiên thử thêm để chắc chắn