Module 2

Reinforcement Learning là gì?

RL là một quá trình dựa trên phản hồi (feedback-based), trong đó tác nhân (agent) tương tác lặp đi lặp lại với môi trường (environment).
Mỗi bước, tác nhân chọn một hành động (action), môi trường phản hồi bằng phần thưởng (reward) và trạng thái mới (state).
Tác nhân học từ chính kinh nghiệm (experience) của mình để cải thiện hiệu suất qua thời gian.
Action-Value function
Giá trị của hành động ( ) là giá trị kỳ vọng của tất cả các giá trị khả thi khi thực hiện hành động a
q_*
$q_*(a) \doteq \mathrm{E}[R_t \mid A_t = a] \; \forall a \in \{1, \ldots, k\}$
$= \sum p(r \mid a)\, r$
Giá trị của hành động là số chưa biết -> cần được ước tính! : giá trị kỳ vọng thực sự : giá trị kỳ vọng ước tính
Mục tiêu là chọn hành động a để tối đa hóa phần thưởng/giá trị kỳ vọng của hành động

$$\arg\max_a q_*(a)$$

Module 3

Chính sách (Policies) :

Một policy trong RL là một bản đồ (mapping) từ mỗi trạng thái sang một phân phối xác suất trên tập các hành động khả dĩ.
$\pi(a \mid s) = P\,(A_t = a \mid S_t = s)$

Hàm giá trị (Value Functions)

Value functions giúp ước lượng mức độ "tốt" khi ở một state (hoặc khi thực hiện một action từ state đó) dưới một policy nhất định.
Dựa vào các giá trị này, agent biết nên ưu tiên vào state nào, action nào để tối đa hóa tổng reward dài hạn
Hai loại value functions chính:
1. State Value Function :
Định nghĩa: kỳ vọng tổng return khi bắt đầu từ trạng thái và theo policy trong tương lai.

Công thức:

$$V^{\pi}(s) = \mathrm{E}_{\pi}[G_t \mid S_t = s] = \mathrm{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s],$$

trong đó:
là tổng discounted return từ thời điểm .
là discount factor .
Kỳ vọng tính trên phân phối hành động do sinh ra và phân phối chuyển tiếp trạng thái.
Ý nghĩa: cho biết "goodness" của state nếu agent cứ tiếp tục theo policy .
2. Action Value Function :
Định nghĩa: kỳ vọng tổng return nếu agent bắt đầu từ trạng thái , thực hiện action ngay, rồi tiếp tục theo policy sau đó.
Công thức:

$$Q^{\pi}(s,a) = \mathrm{E}_{\pi}[G_t \mid S_t = s, A_t = a] = \mathrm{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a].$$

Ý nghĩa: biểu diễn "goodness" của việc chọn khi ở , rồi tiếp tục theo π

Module 4

Chỉ định chính sách(Specifying policies)

Chính sách(policy) là chiến lược được tác nhân sử dụng để quyết định hành động nào sẽ thực hiện trong trạng thái nhất định. Về cơ bản, chính sách

xác định hành vi của tác nhân tại bất kỳ thời điểm nào

Chính sách là hàm ánh xạ trạng thái thành hành động:

• Chính sách xác định: $\pi(s) = a$, trong đó đối với mọi trạng thái, nó trả về một hành động cụ thể a

• Chính sách ngẫu nhiên: $\pi(a|s) = P(a|s)$, phân phối xác suất trên các hành động cho một

trạng thái

Chính sách (Policies)chỉ phụ thuộc vào trạng thái hiện tại, không phụ thuộc vào các yếu tố khác như thời gian hoặc các

trạng thái trước đó.

Generalized Policy Iteration (GPI) là gì?

Khái niệm cốt lõi:

GPI là khung chung hợp nhất nhiều thuật toán trong RL, dựa trên nguyên lý "lặp lại hai bước: ước lượng giá trị → cải tiến chính sách" cho đến khi hội tụ.

Trong mọi thuật toán thuộc GPI, luôn tồn tại hai quá trình song song:

1. Policy Evaluation: Đánh giá (ước lượng) hàm giá trị của chính sách hiện tại.
2. Policy Improvement: Dựa trên giá trị vừa ước lượng, cải tiến (thường là chọn hành động greedy) để sinh ra chính sách mới tốt hơn.

Quá trình này có thể thực hiện cho đến khi không còn cải thiện được nữa, tức khi chính sách đạt tối ưu.

Điểm khác biệt so với Policy Iteration "thuần túy":

Trong Policy Iteration chuẩn, ta thường khóa hoàn toàn ước lượng giá trị cho đến khi hội

tụ "gần đúng" rồi mới làm cải tiến.

Trong GPI, không nhất thiết chờ policy evaluation hoàn chỉnh. Ví dụ: Value Iteration chỉ chạy một lần quét (sweep) để ước lượng giá trị và ngay lập tức "greedify" → lặp lại liên tục.

3. Các bước cơ bản trong GPI

GPI không gán cứng số lượt cho mỗi bước; thay vào đó, hai quá trình đánh giá và cải tiến diễn ra

một cách linh hoạt:

1. Policy Evaluation (Ước lượng giá trị)

Mục đích: Với policy cho trước (có thể là policy được khởi tạo ngẫu nhiên hoặc đã cải

tiến từ bước trước), tính hoặc xấp xỉ hàm giá trị V^{π} (hoặc Q^{π}).

Cách làm: Đầy đủ (Full): Giải hệ Bellman Expectation cho V^{π}
(như trong iterative policy evaluation)

Một phần (Partial): Chỉ cập nhật một số state nhất định (hoặc một quét "non-sweep") rồi dừng, để chuyển sang bước cải tiến ngay.

Ví dụ: Trong Value Iteration, mỗi quét qua toàn bộ state để cập nhật

$$V(s) \leftarrow \max_a \sum_{s',r} P(s', r \mid s, a)[r + \gamma V(s')]$$

chính là kết hợp đồng thời việc tính tiếng nói của policy cải tiến và ước lượng giá trị

2. Policy Improvement (Cải tiến chính sách)

Mục đích: Dựa trên V^π (hoặc Q^π) vừa ước lượng, chọn hành động tốt nhất (greedy) ở mỗi state để tạo ra policy mới .

Công thức mẫu (dùng giá trị trạng thái):

$$\pi'(s) = \arg\max_a \sum_{s',r} P(s', r \mid s, a)[r + \gamma V^\pi(s')].$$

Trường hợp dùng trực tiếp Q^π:

$$\pi'(s) = \arg\max_a Q^\pi(s, a).$$

3. Lặp lại cho đến khi hội tụ
Các bước trên liên tục "chạy song song": không phải đợi evaluation hoàn chỉnh mới cải tiến, mà có thể xen kẽ nhiều lần hơn.
Khi policy không thay đổi qua bước improvement, GPI đã tìm được optimal policy π^* và optimal value function V^* hoặc Q^*

4. Value Iteration như một trường hợp GPI

Khái niệm: Value Iteration thực chất là GPI ở mức "nhanh gọn": mỗi lần chỉ làm một bước cập
nhật giá trị rồi ngay lập tức – tức thì giả định policy mới (greedy) dựa trên giá trị vừa cập nhật,
mà không thực hiện Policy Evaluation đến khi hội tụ.
Công thức chính:

$$V_{k+1}(s) = \max_a \sum_{s',r} P(s', r \mid s, a)[r + \gamma V_k(s')].$$

Ở mỗi bước , với giá trị cuống , ta chọn hành động sao cho kỳ vọng “reward tức thì + chiết khấu giá trị trạng thái kế tiếp” là lớn nhất.
Việc này vừa tương đương với bước Giá trị (Evaluation)—thử nghiệm một giả định policy
greedy—vừa tương đương với bước Cải tiến—cf. greedy action.