2. Mô hình là gì? (What is a Model?)

Trong RL, **mô hình (model)** là một thành phần giả lập cho phép agent **dự đoán** hậu quả của các hành động mà không cần tương tác trực tiếp với môi trường thật.

- **Chức năng cơ bản**: Khi agent ở trạng thái s thực hiện hành động a, mô hình phải cho biết:
  1. **Trạng thái kế tiếp** s' (hoặc phân phối của s').
  2. **Reward** r thu được.
- **Lợi ích chính**:
  - **Planning** (lập kế hoạch): agent sử dụng mô hình để **mô phỏng** ("rollout") nhiều kịch bản hành động – kết quả là có thể cải thiện chính sách (π) mà không cần tốn nhiều tương tác thực.
  - **Tăng hiệu quả mẫu** (sample efficiency): nhờ mô phỏng, agent cần ít tương tác với môi trường thật để đạt được chính sách chất lượng tương tự.

---

3. Phân loại mô hình (Model Types)

3.1. Sample Model

- **Mô tả**: Mỗi lần agent gọi sample model với (s, a), mô hình sẽ **trả về một kết quả cụ thể** (s', r), được sinh ngẫu nhiên theo xác suất thực của môi trường.
- **Tiến hành**: không trả về toàn bộ phân phối, chỉ trả về **một mẫu** duy nhất mỗi lần.
- **Ví dụ đơn giản**:
  - **Lật đồng xu**: sample model sẽ "lật" và trả về "heads" hoặc "tails" với xác suất 50–50 mỗi lần gọi.
  - **Gridworld**: khi bạn gọi sample model ở ô (2,3) với action "phải", mô hình có thể ngẫu nhiên trả về ô (2,4) cùng reward 0, hoặc (3,3) cùng reward -1 nếu có tính lỗi trượt đường.

3.2. Distribution Model

- **Mô tả**: Khi agent hỏi phân phối cho (s, a), model sẽ **trả về toàn bộ phân phối xác suất** $P(s', r \mid s, a)$ – tức xác suất chuyển tới mỗi trạng thái s' và nhận mỗi giá trị reward r.
- **Tiến hành**: cung cấp **bảng xác suất** hoặc hàm phân phối rõ ràng.
- **Ví dụ đơn giản**:
  - **Lật đồng xu**: distribution model cho biết "Heads với xác suất 0.5, Tails với xác suất 0.5".
  - **Gridworld**: distribution model cho biết "Nếu agent ở ô (2,3) và đi phải, thì 80% tới ô (2,4) reward=0, 10% trượt xuống ô (3,3) reward=0, 10% trượt lên ô (1,3) reward=0".

---

4. Khi nào dùng Sample Model hay Distribution Model?

Tình huống / Đặc điểm môi trường	Sample Model	Distribution Model
Môi trường xác định (deterministic)	Thường dùng sample model (dễ triển khai, đơn giản).	Cũng dùng được, nhưng không cần thiết do phân phối đậm đặc.
Môi trường ngẫu nhiên (stochastic)	Dùng sample model khi bạn chỉ cần chạy **mô phỏng** nhiều lần.	Dùng distribution model khi muốn **tính toán chính xác** (ví dụ giải phương trình Bellman) hoặc khi cần phân tích xác suất.
Không gian trạng thái lớn	Sample model gọn nhẹ (không cần lưu cả bảng phân phối).	Distribution model có thể quá cồng kềnh, tốn bộ nhớ.
Cần phân tích lý thuyết / planning	Khó ứng dụng trực tiếp (do chỉ có mẫu).	Rất phù hợp vì có đầy đủ thông tin xác suất để giải Bellman hoặc planning (ví dụ Value Iteration).

5. Ưu/nhược điểm của mỗi loại

Tiêu chí	Sample Model	Distribution Model
Thông tin	Chỉ cung cấp 1 mẫu cụ thể, thiếu tổng quan về phân phối.	Cung cấp đầy đủ phân phối – bao quát mọi khả năng chuyển tiếp và reward.
Bộ nhớ & Tính toán	Nhẹ, chỉ sinh ngẫu nhiên theo quy tắc, tốn ít bộ nhớ.	Nặng, phải lưu hoặc tính toán toàn bộ bảng xác suất, đặc biệt với nhiều trạng thái/ hành động.
Triển khai	Dễ cài đặt: chỉ cần sinh mẫu theo quy tắc.	Phức tạp: phải thiết kế và duy trì phân phối cho từng (s, a).
Ứng dụng	Thích hợp cho **mô phỏng** (MC planning, model-based RL sample-based).	Thích hợp cho **phân tích**, DP planning (Value/Policy Iteration).
Độ chính xác	Biến thiên do mẫu ngẫu nhiên, cần nhiều mẫu để hội tụ.	Chính xác về mặt xác suất, không cần sinh nhiều mẫu để biết phân

Tiêu chí	Sample Model	Distribution Model
		phối.

6. Ví dụ minh họa

6.1. Ví dụ Distribution Model: Two-Armed Bandit

- **Bối cảnh**: bài toán multi-armed bandit với 2 cần gạt (arm). Mỗi arm trả về reward 0 hoặc 1 theo phân phối Bernoulli.
- **Distribution Model**:
  - Arm 1: $P(r=1)=0.7,\ P(r=0)=0.3.$
  - Arm 2: $P(r=1)=0.4,\ P(r=0)=0.6.$
- **Cách dùng**: nếu cần tính chính xác kỳ vọng hoặc phương sai, model cho ta phân phối reward. Planning (VD: tính giá trị kỳ vọng của chiến lược chọn arm 1 liên tục) dễ dàng tính bằng công thức, không cần mô phỏng.

6.2. Ví dụ Sample Model: Simple Gridworld

- **Bối cảnh**: grid 3×3, agent có 4 hành động (↑↓←→). Một số ô có reward cố định (vd: ô đích $+1$, ô bẫy -1). Còn lại reward = 0.
- **Sample Model**:
  - Khi gọi sample_model(state=(1,1), action=→), mô hình dựa trên các xác suất định sẵn (ví dụ trượt 10%) sẽ **trả về** một cặp (s', r) cụ thể, ví dụ $((1,2), 0)$.
  - Để thu thập ước tính giá trị, agent chỉ việc lặp nhiều lần gọi sample_model, rồi dùng Monte Carlo hoặc n-step TD trên các mẫu này.

7. Tóm tắt (Summary)

- **Mô hình (Model)** trong RL cho phép **dự đoán** trạng thái kế tiếp và reward mà không cần tương tác thật.
- **Sample Model**: trả về mẫu cụ thể, triển khai đơn giản, dùng cho **mô phỏng**.
- **Distribution Model**: cho phân phối đầy đủ, dùng cho **planning** chính xác, nhưng cồng kềnh.
- **Lựa chọn**: tùy vào quy mô, tính chất môi trường và mục tiêu (mô phỏng nhanh hay phân tích chính xác)