Tổng quan về Dyna: Kết hợp giữa Học trực tiếp (Direct RL) và Lập kế hoạch (Planning)

1. **Hai nguồn trải nghiệm chính:**
    - **Trải nghiệm thực từ môi trường** (environment experience) — như trong Q-learning, ta học trực tiếp từ các bước đi thực tế.
    - **Trải nghiệm giả lập từ model** (simulated experience from the model) — ta có thể mô phỏng các bước đi dựa trên model đã học để phục vụ cho lập kế hoạch.
2. **Ý tưởng của Dyna:**
    - Dyna là một kiến trúc tổng hợp giữa việc học trực tiếp và lập kế hoạch.
    - Nó dùng cả hai nguồn trải nghiệm: môi trường thật và trải nghiệm mô phỏng từ model để cập nhật giá trị.
    - Model được học dần dần từ trải nghiệm thực.
    - Từ model, ta có thể tạo ra các trải nghiệm giả lập (simulated transitions) để thực hiện planning.
    - Việc chọn trạng thái và hành động để mô phỏng gọi là **search control**.
3. **Ví dụ về robot trong mê cung:**
    - Robot bắt đầu không biết gì, di chuyển ngẫu nhiên đến khi lần đầu gặp mục tiêu.
    - Cập nhật giá trị Q cho hành động có phần thưởng thực (direct RL).
    - Robot học model dựa trên trải nghiệm: biết chuyển từ trạng thái nào sang trạng thái nào với phần thưởng bao nhiêu.
    - Trong các bước tiếp theo, robot dùng model để mô phỏng nhiều lần các bước đi đã từng trải nghiệm, cập nhật giá trị Q qua planning.
    - Nhờ vậy, robot nhanh chóng cải thiện chính sách, tìm đường ngắn hơn tới đích chỉ sau vài lần chạy, trong khi Q-learning thuần túy cần nhiều tập episode hơn.

Tabular Dyna-Q: Một trường hợp cụ thể của Dyna

1. **Model đơn giản giả định chuyển tiếp deterministic (xác định):**
    - Ví dụ con thỏ ở trạng thái A đi phải sang trạng thái B luôn với phần thưởng 0.
    - Model ghi nhớ cặp (state, action) và kết quả (next state, reward).
    - Mỗi trải nghiệm từ môi trường giúp xây dựng model chính xác dần.
2. **Quy trình Dyna-Q:**
    - **Tương tác với môi trường:** Chọn hành động theo chính sách epsilon-greedy, quan sát phần thưởng và trạng thái kế tiếp.
    - **Cập nhật trực tiếp (Direct RL):** Dùng Q-learning để cập nhật giá trị Q từ trải nghiệm thật.
    - **Học model:** Lưu lại kết quả chuyển tiếp (state, action → next state, reward).
    - **Planning:** Thực hiện nhiều lần các bước sau:
        - **Search control:** Chọn ngẫu nhiên một cặp (state, action) đã trải nghiệm.
        - **Model query:** Dự đoán trạng thái kế tiếp và phần thưởng từ model.
        - **Value update:** Thực hiện cập nhật Q-learning trên dữ liệu giả lập đó.
3. **Kết quả và minh họa:**
    - Robot ban đầu mất nhiều bước đi khi chưa có model.
    - Sau khi có model và thực hiện nhiều bước planning trên mỗi bước môi trường, giá trị Q nhanh chóng lan truyền trên không gian trạng thái.
    - Robot có thể tìm đường ngắn hơn (ví dụ giảm từ 184 bước xuống còn 18 bước chỉ sau 2 episode).
    - Dyna-Q tận dụng trải nghiệm hiệu quả hơn nhờ lập kế hoạch mô phỏng, không chỉ học từ trải nghiệm thực.