1. Định nghĩa Temporal Difference Learning (TD learning)

- TD learning là một phương pháp học giá trị (value function) trong reinforcement learning.
- Mục tiêu là ước lượng giá trị của các trạng thái (state values) dựa trên các trải nghiệm (trajectories) trong môi trường.
- TD learning kết hợp các ưu điểm của Monte Carlo (MC) và Dynamic Programming (DP).
- Khác với MC phải đợi đến cuối episode mới cập nhật giá trị, TD cập nhật giá trị ngay trong quá trình thu thập dữ liệu (incremental update).

2. Sự khác biệt so với Monte Carlo:

- MC cập nhật giá trị dựa trên tổng phần thưởng (returns) của toàn bộ episode, phải đợi tới cuối episode mới tính được.
- TD thay thế giá trị tổng returns bằng một **ước lượng bootstrapped**: sử dụng phần thưởng ở bước hiện tại cộng với giá trị ước lượng của trạng thái kế tiếp.

3. Công thức cập nhật TD:

Giá trị của trạng thái ở thời điểm t, V(St), được cập nhật về gần giá trị:

$$R_{t+1} + \gamma V(S_{t+1})$$

trong đó:

- R_{t+1} là phần thưởng nhận được sau khi chuyển từ trạng thái S_t sang S_{t+1}
- γ là hệ số giảm giá (discount factor)
- $V(S_{t+1})$ là giá trị ước lượng của trạng thái kế tiếp

4. TD Error (Sai số TD), ký hiệu δ_t:

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

- Sai số này đo khoảng cách giữa giá trị hiện tại của trạng thái và mục tiêu cập nhật bootstrapped.
- Đây chính là "error" dùng để điều chỉnh giá trị ước lượng $V(S_t)$.

---

5. Thuật toán TD(0):

- Khởi tạo giá trị ước lượng V cho tất cả các trạng thái.
- Trong mỗi bước thời gian:
  - Quan sát trạng thái S_t, hành động, phần thưởng R_{t+1}, trạng thái kế tiếp S_{t+1}.
  - Cập nhật $V(S_t)$ theo công thức:

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t = V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$$

với α là tốc độ học (learning rate).

- Không cần phải lưu lại toàn bộ episode, chỉ cần lưu trạng thái trước đó để cập nhật.

6. So sánh với Dynamic Programming (DP):

- DP cần mô hình môi trường để tính giá trị kỳ vọng trên tất cả các trạng thái kế tiếp.
- TD chỉ cần quan sát trạng thái kế tiếp thực tế từ môi trường, không cần mô hình.

---

7. Ý nghĩa:

- TD learning giúp học giá trị một cách hiệu quả, có thể học ngay trong quá trình tương tác với môi trường.
- Là nền tảng cho nhiều thuật toán quan trọng trong Reinforcement Learning.

Nếu bạn muốn, tôi có thể giúp bạn:

- Giải thích chi tiết hơn về TD error và cách nó ảnh hưởng đến cập nhật giá trị.

- Minh họa bằng ví dụ cụ thể.
- So sánh trực quan TD với Monte Carlo và Dynamic Programming.