Generalized Policy Iteration (GPI)

Là khung tổng quát gồm hai bước lặp lại liên tục:

1. **Policy Evaluation**: Đánh giá chính sách hiện tại.
2. **Policy Improvement**: Cải thiện chính sách dựa trên giá trị đã đánh giá.

TD Learning (Temporal Difference Learning)

Là phương pháp học dựa trên sự khác biệt tạm thời giữa giá trị ước lượng hiện tại và giá trị quan sát được từ môi trường. TD cập nhật giá trị **từng bước** thay vì đợi đến cuối episode.

Sarsa – *On-policy TD Control Algorithm*

- Viết tắt của: **State, Action, Reward, next State, next Action**.
- Dùng để **ước lượng giá trị hành động** Q(s,a)Q(s, a)Q(s,a) theo chính sách hiện tại.

- Công thức cập nhật:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma Q(s',a') - Q(s,a) \right]$$

- **On-policy**: Hành động tiếp theo a' được chọn từ chính sách đang học (thường là ε-greedy).

Q-learning – *Off-policy TD Control Algorithm*

- Là thuật toán học tăng cường **phổ biến nhất**, học trực tiếp giá trị tối ưu.
- Cập nhật theo công thức:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

- **Off-policy**: Luôn chọn hành động tối ưu (greedy) trong cập nhật, bất kể chính sách hiện tại.