Multi-Step TD Target

Shusen Wang

Sarsa versus **Q-Learning**

- **Sarsa** is for training action-value function, $Q_\pi(s, a)$.
- TD target: $y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1})$.

Sarsa versus Q-Learning

- **Sarsa** is for training action-value function, $Q_\pi(s, a)$.
- TD target: $y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1})$.

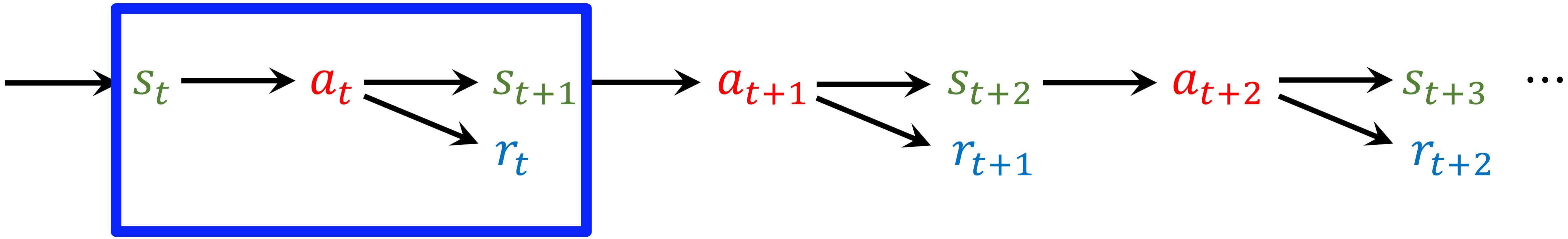
- **Q-learning** is for training the optimal action-value function, $Q^\star(s, a)$.

Sarsa versus Q-Learning

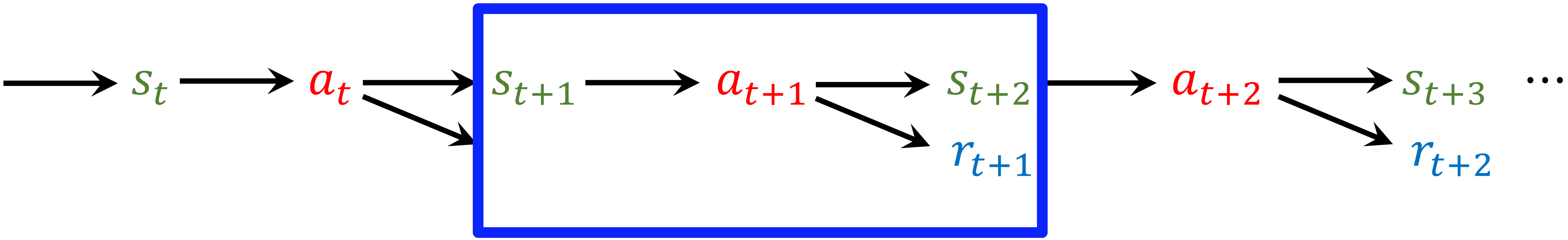
- **Sarsa** is for training action-value function, $Q_\pi(s, a)$.
- TD target: $y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1})$.

- **Q-learning** is for training the optimal action-value function, $Q^\star(s, a)$.
- TD target: $y_t = r_t + \gamma \cdot \max_a Q^\star(s_{t+1}, a)$.

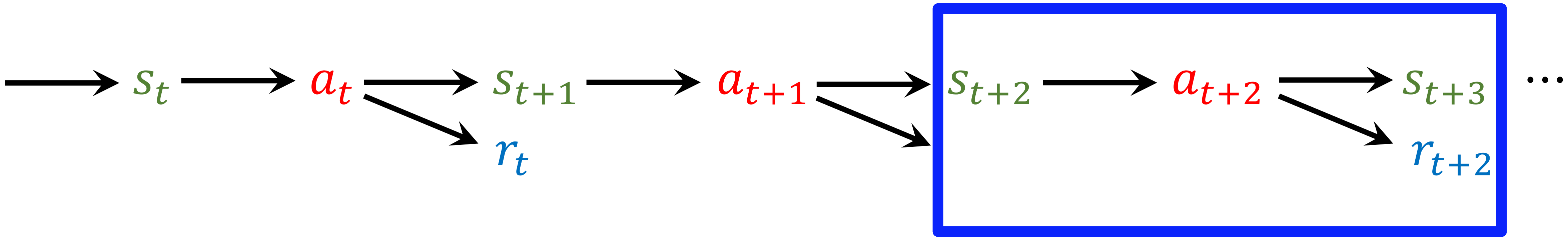
Using One Reward



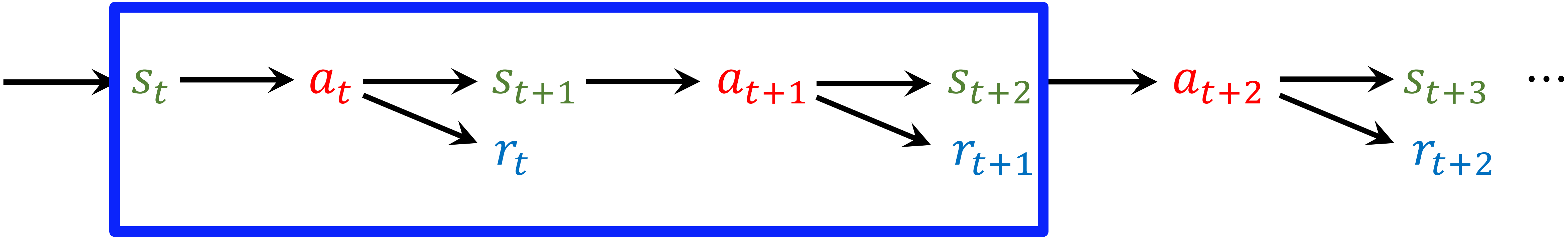
Using One Reward



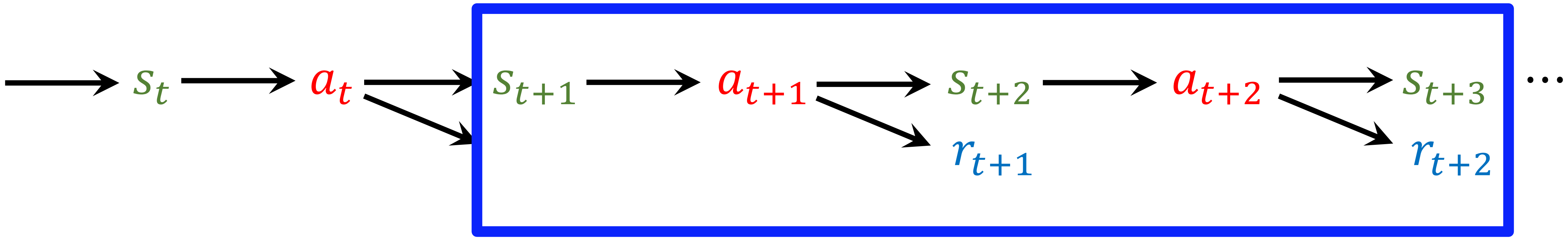
Using One Reward



Using Multiple Rewards



Using Multiple Rewards



Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}.$

Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

$= R_{t+1} + \gamma \cdot U_{t+2}$

Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}.$

$$= R_{t+1} + \gamma \cdot U_{t+2}$$

Identity: $U_t = R_t + \gamma \cdot \left(R_{t+1} + \gamma \cdot U_{t+2}\right).$

Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}.$

$$= R_{t+1} + \gamma \cdot U_{t+2}$$

Identity: $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot U_{t+2}.$

Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}.$

Identity: $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot U_{t+2}.$

Identity: $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot U_{t+3}.$

Multi-Step Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

Identity: $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot U_{t+2}$.

Identity: $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot U_{t+3}$.

Identity: $U_t = \sum_{i=0}^{m-1} \gamma^i \cdot R_{t+i} + \gamma^m \cdot U_{t+m}$.

Multi-Step TD Targets

Identity: $U_t = \sum_{i=0}^{m-1} \gamma^i \cdot R_{t+i} + \gamma^m \cdot U_{t+m}.$

- m-step TD target for **Sarsa**:

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot Q_\pi(s_{t+m}, a_{t+m}).$$

Multi-Step TD Targets

Identity: $U_t = \sum_{i=0}^{m-1} \gamma^i \cdot R_{t+i} + \gamma^m \cdot U_{t+m}$.

- m-step TD target for **Sarsa**:

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot Q_\pi(s_{t+m}, a_{t+m}).$$

- One-step TD target for **Sarsa**:

$$y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1}).$$

Multi-Step TD Targets

Identity: $U_t = \sum_{i=0}^{m-1} \gamma^i \cdot R_{t+i} + \gamma^m \cdot U_{t+m}.$

- m-step TD target for **Q-learning**:

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot \max_a Q^\star(s_{t+m}, a).$$

Multi-Step TD Targets

Identity: $U_t = \sum_{i=0}^{m-1} \gamma^i \cdot R_{t+i} + \gamma^m \cdot U_{t+m}.$

- m-step TD target for **Q-learning**:

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot \max_a Q^\star(s_{t+m}, a).$$

- One-step TD target for **Q-learning**:

$$y_t = r_t + \gamma \cdot \max_a Q^\star(s_{t+1}, a).$$

One-Step versus Multi-Step

- One-step TD target uses only one reward: r_t.
- m-step TD target uses m rewards: $r_t, r_{t+1}, r_{t+2}, \cdots, r_{t+m-1}$.
- If m is suitably tuned, m-step target works better than one-step target [1].

Reference:

1. Hossel et al. Rainbow: combining improvements in deep reinforcement learning. In *AAAI*, 2018.

Thank you!