

Temporal-difference Learning

(MC ideas + DP ideas)

↑
learn from
experience

↑
update estimates
based in part on other learned
estimates, without waiting
for a final outcome

{ prediction: estimate value function V_π for a given policy π
control: find an optimal policy

MC / constant- α MC:

$$V(S_t) \leftarrow V(S_t) + \alpha [G_t - V(S_t)]$$

TD(0) / one-step TD:

$$V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

Full episode (green arrow pointing to G_t)

Sample (green arrow pointing to $R_{t+1} + \gamma V(S_{t+1})$)

Next time step (green arrow pointing to S_{t+1})

Current estimate (green arrow pointing to $V(S_t)$)

$$V_\pi(s) = \mathbb{E}_\pi [G_t | S_t = s] \quad (\text{MC})$$

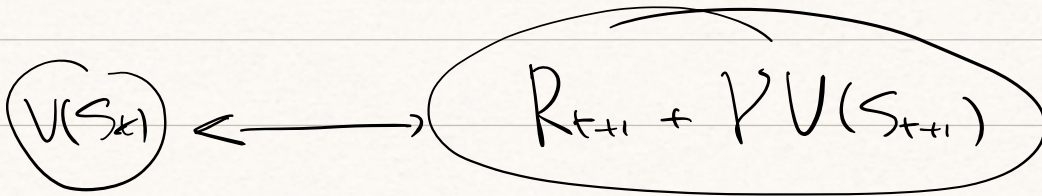
$$= \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$$= \mathbb{E}_\pi [R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s] \quad (\text{TD}(0))$$

"sample updates"

looking ahead to a sample successor state
(or state-action pair)

TD error :



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

MC error :

$$G_t - V(S_t) = \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k$$

TD vs DP

DON'T require a model

TD vs MC ← Wait until the end of an episode

consider transitions from state-action pair to state-action pair, and learn the values of

Sarsa: On-policy TD control state-action pairs.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

$(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$

2 for:

$$TD: V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

$$Sarsa: Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

Q-learning: Off-policy TD control

Watkins, 1989

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Expected Sarsa

$$\begin{aligned} Q(S_t, A_t) &\leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \mathbb{E}_{\pi} [Q(S_{t+1}, A_{t+1}) | S_{t+1}] - Q(S_t, A_t)] \\ &\leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \sum_a \pi(a | S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)] \end{aligned}$$

Double Learning

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) +$$

$$\alpha [R_{t+1} + \gamma Q_2(S_{t+1}, \arg \max_a Q_1(S_{t+1}, a)) - Q_1(S_t, A_t)]$$