Temporal Difference Learning

Shangtong Zhang

University of Virginia

Difficulties in implementing dynamic programming

- Unknown reward function
- Unknown transition function

From DP to TD

- Full update to incremental update
- Synchronous update to asynchronous update
- Deterministic update to stochastic update

Discounted total rewards – prediction

$$
\begin{aligned}
v_{t+1} &\leftarrow \mathcal{T}_\pi v_t \\
v_{t+1}(s) &\leftarrow \begin{cases} v_t(s) + \alpha_t \left(R_{t+1} + \gamma v_t(S_{t+1}) - v_t(s)\right), & s = S_t \\ v_t(s), & s \neq S_t \end{cases}
\end{aligned}
$$

Multistep TD

$$\begin{aligned} v_{t+1} &\leftarrow \mathcal{T}_{\pi}^{(2)} v_t \\ \delta_t &\leftarrow R_{t+1} + \gamma R_{t+2} + \gamma^2 v_t(S_{t+2}) - v_t(S_t) \\ v_{t+1}(s) &\leftarrow \begin{cases} v_t(s) + \alpha_t \delta_t, & s = S_t \\ v_t(s), & s \neq S_t \end{cases} \end{aligned}$$

Monte Carlo

$$\begin{aligned} v_{t+1} &\leftarrow \mathcal{T}_\pi^{(T-1)} v_t \\ \delta_t &\leftarrow R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T - v_t(S_t) \\ v_{t+1}(s) &\leftarrow \begin{cases} v_t(s) + \alpha_t \delta_t, & s = S_t \\ v_t(s), & s \neq S_t \end{cases} \end{aligned}$$

Supervised Learning v.s. Reinforcement Learning

$$\begin{aligned} & v_\pi(s) \\ =& \mathbb{E}\left[G_t | S_t = s\right] \\ =& \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T | S_t = s\right] \end{aligned}$$

- SL: known label
- RL: guessed "label" (bootstrapping)

- Monte Carlo: SL and gradient descent
- TD: RL and <u>semi-gradient</u> descent

Discounted total rewards – prediction

$$
\begin{aligned}
q_{t+1} &\leftarrow \mathcal{T}_\pi q_t \\
\delta_t &\leftarrow R_{t+1} + \gamma q_t(S_{t+1}, A_{t+1}) - q_t(S_t, A_t) \\
q_{t+1}(s, a) &\leftarrow \begin{cases} q_t(s, a) + \alpha_t \delta_t, & (s, a) = (S_t, A_t) \\ q_t(s, a), & (s, a) \neq (S_t, A_t) \end{cases}
\end{aligned}
$$

Discounted total rewards – off-policy prediction

$$\begin{aligned} v_{t+1} &\leftarrow \mathcal{T}_\pi v_t \\ v_{t+1}(s) &\leftarrow \begin{cases} v_t(s) + \alpha_t \left(\rho_t \left(R_{t+1} + \gamma v_t(S_{t+1})\right) - v_t(s)\right), & s = S_t \\ v_t(s), & s \neq S_t \end{cases} \end{aligned}$$

$$v_{t+1}(s) \leftarrow \begin{cases} v_t(s) + \alpha_t \rho_t \left(R_{t+1} + \gamma v_t(S_{t+1}) - v_t(s)\right), & s = S_t \\ v_t(s), & s \neq S_t \end{cases}$$

Discounted total rewards – off-policy prediction

$$\begin{aligned}
q_{t+1} \leftarrow & \mathcal{T}_\pi q_t \\
\delta_t \leftarrow & R_{t+1} + \gamma \rho_{t+1} q_t(S_{t+1}, A_{t+1}) - q_t(S_t, A_t) \\
\delta_t \leftarrow & R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) q_t(S_{t+1}, a) - q_t(S_t, A_t) \\
q_{t+1}(s, a) \leftarrow & \begin{cases} q_t(s, a) + \alpha_t \delta_t, & (s, a) = (S_t, A_t) \\ q_t(s, a), & (s, a) \neq (S_t, A_t) \end{cases}
\end{aligned}$$

Multistep off-policy prediction

$$
\begin{aligned}
v_{t+1} &\leftarrow \mathcal{T}_\pi^{(2)} v_t \\
G_{t:t+2} &\leftarrow \rho_t \rho_{t+1} \left(R_{t+1} + \gamma R_{t+2} + \gamma^2 v_t(S_{t+2}) \right) \\
G_{t:t+2} &\leftarrow \rho_t R_{t+1} + \rho_t \rho_{t+1} \left(\gamma R_{t+2} + \gamma^2 v_t(S_{t+2}) \right) \\
v_{t+1}(s) &\leftarrow \begin{cases} v_t(s) + \alpha_t \left(G_{t:t+2} - v(s) \right), & s = S_t \\ v_t(s), & s \neq S_t \end{cases}
\end{aligned}
$$

Multistep off-policy prediction

$$\begin{aligned}
q_{t+1} &\leftarrow \mathcal{T}_\pi^{(2)} q_t \\
G_{t:t+2} &\leftarrow \rho_{t+1}\rho_{t+2}\left(R_{t+1} + \gamma R_{t+2} + \gamma^2 q_t(S_{t+2}, A_{t+2})\right) \\
G_{t:t+2} &\leftarrow R_{t+1} + \rho_{t+1}\gamma R_{t+2} + \rho_{t+1}\rho_{t+2}\gamma^2 q_t(S_{t+2}, A_{t+2}) \\
q_{t+1}(s,a) &\leftarrow \begin{cases} q_t(s,a) + \alpha_t\left(G_{t:t+2} - q_t(s,a)\right), & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases}
\end{aligned}$$

Multistep off-policy prediction

$$G_{t:t+2} \leftarrow R_{t+1} + \rho_{t+1}\gamma R_{t+2} + \rho_{t+1}\gamma^2 \sum_a \pi(a|S_{t+2}) q_t(S_{t+2}, a)$$

$$\begin{aligned} G_{t:t+2} \leftarrow & R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) q_t(S_{t+1}, a) \\ & + \gamma \rho_{t+1} \left(R_{t+2} + \gamma \sum_a \pi(a|S_{t+2}) q_t(S_{t+2}, a) \right) \end{aligned}$$

Discounted total rewards – control

Estimating action value function or state value function?

Discounted total rewards – on-policy control

$$\begin{aligned} q_{t+1} &\leftarrow \mathcal{T}_{\pi_{q_t}} q_t \\ \delta_t &\leftarrow R_{t+1} + \gamma q_t(S_{t+1}, A_{t+1}) - q_t(S_t, A_t) \\ q_{t+1}(s,a) &\leftarrow \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases} \end{aligned}$$

Exploration and exploitation dilemma

Discounted total rewards – off-policy control

$$
\begin{aligned}
q_{t+1} &\leftarrow \mathcal{T}_{*} q_t \\
\delta_t &\leftarrow R_{t+1} + \gamma \max_a q_t(S_{t+1}, a) - q_t(S_t, A_t) \\
q_{t+1}(s, a) &\leftarrow \begin{cases} q_t(s, a) + \alpha_t \delta_t, & (s, a) = (S_t, A_t) \\ q_t(s, a), & (s, a) \neq (S_t, A_t) \end{cases}
\end{aligned}
$$

Cliff walking – comparing on- and off-policy control

Example 6.6 in Sutton & Barto's book.

Off-policy expected SARSA and Q-learning

$$
\begin{aligned}
q_{t+1} &\leftarrow \mathcal{T}_{\pi_{q_t}} q_t \\
\delta_t &\leftarrow R_{t+1} + \gamma \sum_a \pi_{q_t}(a|S_{t+1}) q_t(S_{t+1}, a) - q_t(S_t, A_t)
\end{aligned}
$$

$$
q_{t+1}(s,a) \leftarrow \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases}
$$

Is there multistep Q-learning?

Average reward – prediction

$$
\begin{aligned}
\delta_t \leftarrow & R_{t+1} - J_t + \gamma v_t(S_{t+1}) - v_t(S_t) \\
v_{t+1}(s) \leftarrow & \begin{cases} v_t(s) + \alpha_t \delta_t, & s = S_t \\ v_t(s), & s \neq S_t \end{cases} \\
J_{t+1} \leftarrow & J_t + \alpha_t \left(R_{t+1} - J_t \right)
\end{aligned}
$$

Average reward – off-policy prediction

How to estimate $\bar{J}_\pi$?

$$\begin{aligned} \bar{v}_\pi &= r_\pi - \bar{J}_\pi 1 + P_\pi \bar{v}_\pi \\ \implies \bar{J}_\pi &= d^\top \left(r_\pi + P_\pi \bar{v}_\pi - \bar{v}_\pi \right) \end{aligned}$$

Average reward – off-policy prediction

$$\begin{aligned}
\delta_t \leftarrow & R_{t+1} + \gamma v_t(S_{t+1}) - v_t(S_t) \\
v_{t+1}(s) \leftarrow & \begin{cases} v_t(s) + \alpha_t \rho_t \left(\delta_t - J_t\right), & s = S_t \\ v_t(s), & s \neq S_t \end{cases} \\
J_{t+1} \leftarrow & J_t + \alpha_t \rho_t \left(\delta_t - J_t\right)
\end{aligned}$$

Average reward – off-policy control

$$
\begin{aligned}
\delta_t \leftarrow & R_{t+1} + \gamma \max_a q_t(S_{t+1}, a) - q_t(S_t, A_t) \\
q_{t+1}(s, a) \leftarrow & \begin{cases} q_t(s, a) + \alpha_t \left(\delta_t - J_t\right), & (s, a) = (S_t, A_t) \\ q_t(s, a), & (s, a) \neq (S_t, A_t) \end{cases} \\
J_{t+1} \leftarrow & J_t + \alpha_t \left(\delta_t - J_t\right)
\end{aligned}
$$

References

- Reinforcement Learning: An Introduction by Richard Sutton and Andrew Barto
- Learning and Planning in Average-Reward Markov Decision Processes by Yi Wan, Abhishek Naik, and Richard Sutton