1 Temporal-Difference Learning

TD combines some of the features of both MC and DP. TD does not require a model and can learn from interactions (like MC), and TD can bootstrap, thus learn online (without waiting till the end of episodes) (like DP).

TD(λ) unifies DP, MC, TD.

Prediction problem = policy evaluation (i.e., estimating v_π for a given π)

Control problem = finding an optimal policy

1.1 TD Prediction

Definition 1.1 ***constant-α MC**:*

$$V(S_t) \leftarrow V(S_t) + \alpha \underbrace{\left[G_t - V(S_t)\right]}_{\textit{MC error}}$$

Unlike MC which has to wait until the end of an episode to update, TD only has to wait one time step.

Definition 1.2 ***TD(0)**, or one-step TD:*

$$V(S_t) \leftarrow V(S_t) + \alpha \underbrace{\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]}_{\textit{TD error}}$$

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated

Algorithm parameter: step size $\alpha \in (0, 1]$

Init $V(s)$ $\forall s \in S+$, $V(term.) = 0$

Loop for each episode:

  Init S

  Loop for each step of episode:

    $A \leftarrow$ action given by π for S

    Take action A, observe R, S'

    $V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$

    $S \leftarrow S'$

  until S is terminal

- MC target is an estimate, because a sample return is needed for real expected return
- DP target is an estimate, because the next state value is unknown and the current estimate is used
- TD target is an estimate because (1) samples the expected values; (2) and uses current estimate V instead of the true v_π

TD methods combine the sampling of MC with the bootstrapping of DP. TD and MC updates are *sample update* because they involve looking ahead to a sample successor state.

- sample update: based on a single sample sample successor
- expected update: a complete distribution of all possible successors

Definition 1.3 ***TD error***:

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

MC error can be written as a sum of TD errors (think of it this way, TD updates immediately at each time step, MC updates after an episode (which include a lot of timesteps)

$$
\begin{aligned}
\text{MC error} = G_t - V(S_t) &= \underbrace{R_{t+1} + \gamma G_{t+1}}_{G_t} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma (G_{t+1} - V(S_{t+1})) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 (G_{t+2} - V(S_{t+2}) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k
\end{aligned}
$$

1.2 Advantages of TD Prediction Methods

1.3 Optimality of TD(0)

1.4 Learning Objectives (UA RL MOOC

Lesson 1: Introduction to Temporal Difference Learning

1. Define temporal-difference learning

2. Define the temporal-difference error

3. Understand the TD(0) algorithm

Lesson 2: Advantages of TD

4. Understand the benefits of learning online with TD

5. Identify key advantages of TD methods over Dynamic Programming and Monte Carlo methods

6. Identify the empirical benefits of TD learning