[S&B Book] Chapter 6: Temporal-Difference Learning

Tags	

- Temporal-Difference (TD) learning is a combination of Monte Carlo ideas and dynamic programming (DP) ideas
    - Like MC, TD can learn directly from raw experience without a model of the environment's dynamics;
    - Like DP, TD updates estimates based in part on other learned estimates without waiting for a final outcome (they bootstrap).

▼ 6.1 TD Prediction

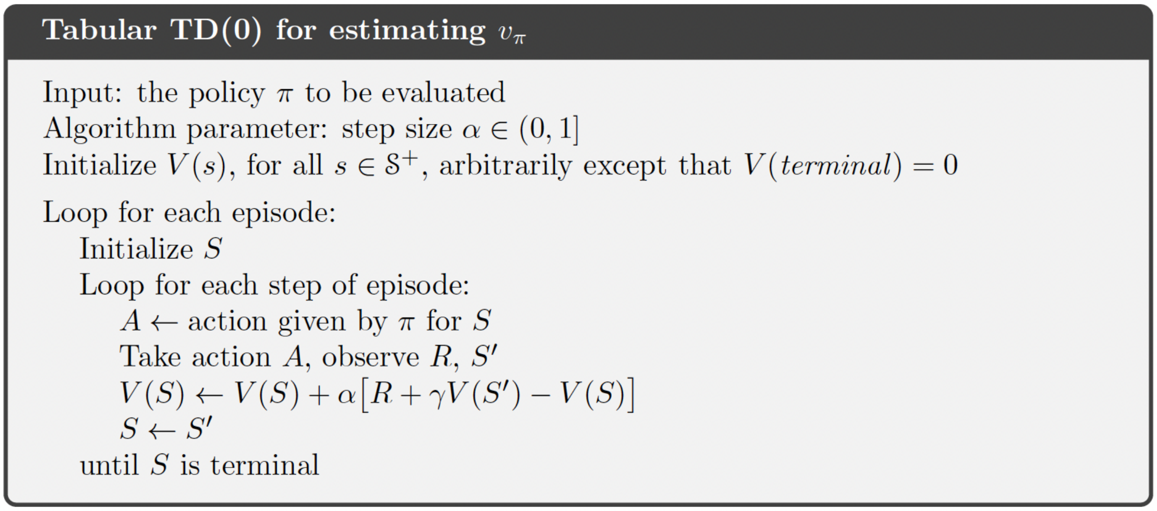
- TD method:

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

  The simplest TD method makes the update immediately on the transition to S_{t+1} and receiving R_{t+1};

  This method is also called ***one-step* TD, or TD(0)**, because it is a special case of the TD(λ) and n-step TD methods.

- Algorithm:



```
Tabular TD(0) for estimating $v_\pi$

Input: the policy $\pi$ to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop for each episode:
    Initialize $S$
    Loop for each step of episode:
        $A \leftarrow$ action given by $\pi$ for $S$
        Take action $A$, observe $R$, $S'$
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until $S$ is terminal
```

- *sample updates* and *expected updates*
  - TD and MC are ***sample updates*** because they involve looking ahead to a sample successor state (or state-action pair), using the value of successor and the reward along the way to compute a back-up value, and then updating the value of the original state (or state-action pair) accordingly;
  - DP methods are ***expected updates*** because they are based on complete distribution of all possible successors;
- **TD error:**

  The difference between the estimated value of S_t and the better estimate $R_{t+1} + \gamma V(S_{t+1})$; the TD error arises in various forms throughout reinforcement learning;

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

- If the array V does not change during the episode (as it does not change in Monte Carlo methods), then the Monte Carlo error can be written as a sum of TD errors:

$$\begin{aligned} G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S) + \gamma V(S_{t+1}) + -\gamma V(S_{t+1}) \\ &= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\ &= \delta_t + \gamma \delta_{t+1} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(G_T - V(S_T)) \\ &= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k \end{aligned}$$

Exercise 6.1:

If V changes during the episode, then (6.6) only holds approximately; what would the difference be between the two sides? Let Vt denote the array of state values used at time t in the TD error (6.5) and in the TD update (6.2). Redo the derivation above to determine the additional amount that must be added to the sum of TD errors in order to equal the Monte Carlo error.

Solution:

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

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

$$\begin{aligned} G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S) + \gamma V(S_{t+1}) + \gamma\alpha\delta_{t+1} - \gamma V(S_{t+1}) - \gamma\alpha\delta_{t+1} \\ &= \delta_t + \gamma\alpha\delta_{t+1} + \gamma(G_{t+1} - [V(S_{t+1}) + \gamma\alpha\delta_{t+1}]) \\ &= \delta_t + \gamma\alpha\delta_{t+1} + \gamma\delta_{t+1} + \gamma^2\alpha\delta_{t+1} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\alpha\delta_T + \gamma^{T-t}(G_T - V(S_T)) \\ &= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k + \gamma^{k-t+1}\alpha\delta_{k+1} \end{aligned}$$

▼ 6.2 Advantages of TD Prediction Methods

Example 6.2

Implementation