Temporal Difference Methods

for Prediction

Rupam Mahmood

February 26, 2020

Worksheet questions:

(*Exercise 6.1 S&B*) If V changes during the episode, then

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

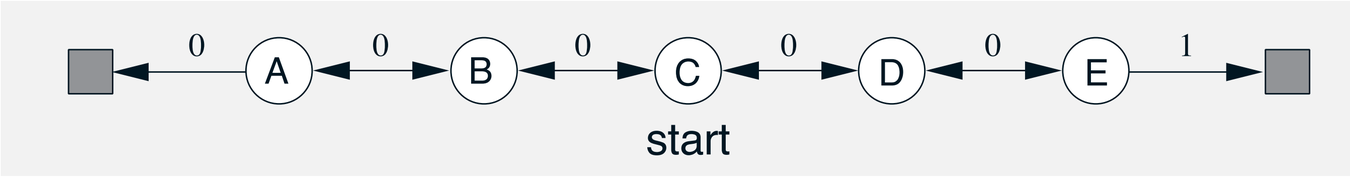
only holds approximately; what would the difference be between the two sides? Let V_t denote the array of state values used at time t in the TD error and in the TD update. Redo the derivation to determine the additional amount that must be added to the sum of TD errors in order to equal the Monte Carlo error.

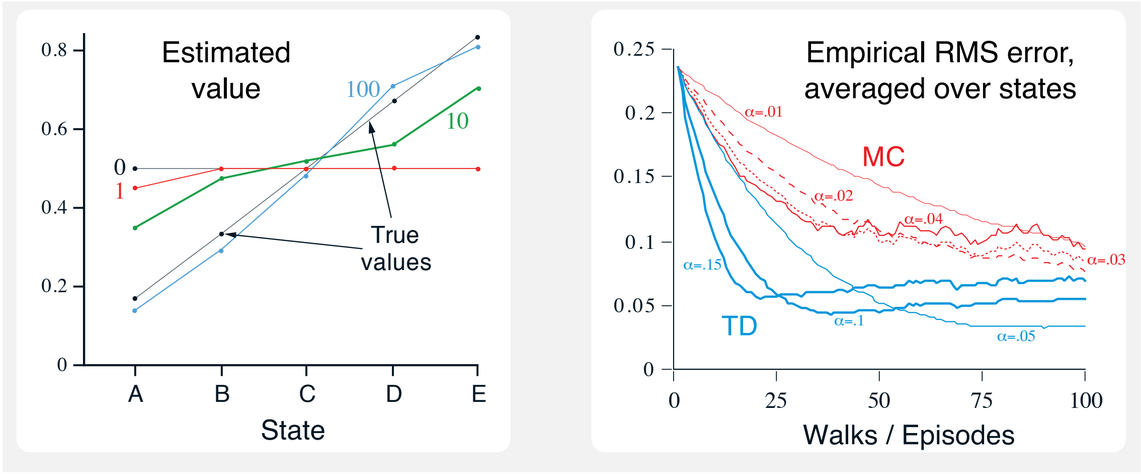
TD(0): $V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]$

Equivalently: $V_{t+1}(S_t) \doteq V_t(S_t) + \alpha \underbrace{\left[R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)\right]}_{\delta_t}$

$$V_{t+1}(s) \doteq V_t(s), \forall s \neq S_t$$

(*Exercise 6.3 S&B*) From the results shown in the left graph of the random walk example it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?





(*Exercise 6.7 S&B*) Design an off-policy version of the TD(0) update that can be used with arbitrary target policy π and covering behavior policy b, using at each step t the importance sampling ratio $\rho_{t:t}$ (5.3).

$$\rho_{t:T-1} \doteq \frac{P(A_t, S_{t+1}, A_{t+1}, \cdots, S_T \,|\, S_t, A_{t:T-1} \sim \pi)}{P(A_t, S_{t+1}, A_{t+1}, \cdots, S_T \,|\, S_t, A_{t:T-1} \sim b)} = \frac{\prod_{k=t}^{T-1} \pi(A_k \,|\, S_k)}{\prod_{k=t}^{T-1} b(A_k \,|\, S_k)}$$

Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π.

Tabular TD(0) for estimating v_π

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

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