Temporal Difference Methods for Prediction

Rupam Mahmood

February 28, 2020

(*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).

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

on-policy sample-average MC: $A_t \sim \pi \implies E_\pi\left[G_t \mid S_t = s\right] = v_\pi(s) \ \rightarrow V(s) \doteq \dfrac{\sum_{t \in \mathcal{T}(s)} G_t}{|\mathcal{T}(s)|}$

off-policy sample-average MC: $A_t \sim b \implies E_b\left[\rho_{t:T-1} G_t \mid S_t = s\right] = v_\pi(s) \ \rightarrow V(s) \doteq \dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T-1} G_t}{|\mathcal{T}(s)|}$

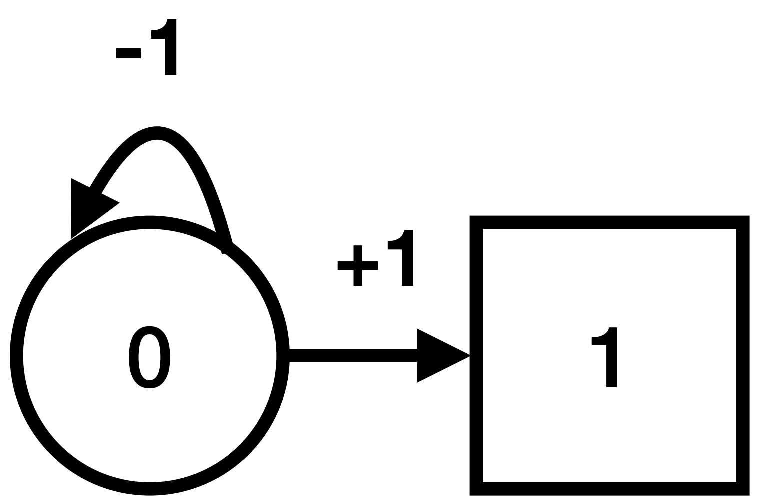
On-policy constant-α MC: $V^{MC}(S_t) \leftarrow V^{MC}(S_t) + \alpha\left[G_t - V^{MC}(S_t)\right]$

Off-policy constant-α MC: ?

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

Off-policy TD(0): ?

Live demo of TD updates



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