September 7, 2022

6 Exercises

6.1 Question

If V changes during the episode, then (6.6) 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 (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.

Answer

If V is updated during the episode, which is the case for TD algorithm, it only makes difference if a state is visited more than once in an episode.

Equation (6.2) becomes:

$V_{t+1}(S_t) = V_t(S_t) + \alpha[R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)]$

$V_{t+1}(S_t) = V_t(S_t) + \alpha\delta_t$

Equation (6.5) becomes:

$\delta_t = R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)$

Monte Carlo error is:

$G_t - V(S_t) = R_{t+1} + \gamma G_{t+1} - V_t(S_t) = \delta_t + \gamma(G_{t+1} - V_t(S_{t+1}))$

With updated values:

$G_t - V(S_t) = \delta_t + \gamma(G_{t+1} - V_{t+1}(S_{t+1}) + \alpha\delta_{t-a})$ where $a \not< 0$ indicates when the last update was.

Note that if a state is visited only once then updated value is never used thus wen say that V_t is similar to V_{t+1}.

For simplicity lets assume a state can only be revisited just after visiting the state. We can introduce an indicator function denoted by f which returns 0 if a state is visited only once and 1 if it is visited again in next step.

$G_t - V(S_t) = \delta_t + \gamma(G_{t+1} - V_{t+1}(S_{t+1}) + \alpha\delta_t f(S_t))$

Continue with the derivation:

$G_t - V(S_t) = \delta_t + \gamma\alpha\delta_t f(S_t) + \gamma(G_{t+1} - V_{t+1}(S_{t+1}))$

$G_t - V(S_t) = \delta_t + \gamma\alpha\delta_t f(S_t) + \gamma(\delta_{t+1} + \gamma\alpha\delta_{t+1} f(S_{t+1}) + \gamma(G_{t+2} - V_{t+2}(S_{t+2})))$

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