VARIANCE OF MC VERSUS TD

ASHWIN RAO

Assume a state is visited only once in an episode, so each of MC and TD update the value function for a state only once in an episode. Assume that for a given episode k, the variance of the value function for all states is the same. We will denote this variance as y_k for MC and z_k for TD. Assume $\gamma = 1$. Assume the variance of observed reward in any episode is a stationary value x. Assume we have n time steps per episode (so, the variance of observed return in any episode is nx).

The MC update is:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t)) = (1-\alpha)V(S_t) + \alpha G_t$$

So, the Variance y_k for MC in episode k is given by:

$$y_k = (1-\alpha)^2 y_{k-1} + \alpha^2 nx = (1-\alpha)^2((1-\alpha)^2 y_{k-2} + \alpha^2 nx) + \alpha^2 nx = \dots$$
$$\dots = (1-\alpha)^{2i} y_{k-i} + \alpha^2 nx \sum_{j=0}^{i-1} (1-\alpha)^{2j}$$

When $i = k, y_{k-i} = y_0 = 0$, and so,

$$y_k = \alpha^2 nx \sum_{j=0}^{k-1} (1-\alpha)^{2j}$$

For large enough k,

$$y_k = \frac{\alpha^2 nx}{1-(1-\alpha)^2} = \frac{\alpha nx}{2-\alpha}$$

The TD update is:

$$V(S_t) \leftarrow V(S_t) + \alpha(R_t + V(S_{t+1}) - V(S_t)) = (1-\alpha)V(S_t) + \alpha(R_t + V(S_{t+1}))$$

So, the Variance z_k for TD in episode k is given by:

$$z_k = (1-\alpha)^2 z_{k-1} + \alpha^2(x + z_{k-1}) = ((1-\alpha)^2 + \alpha^2) z_{k-1} + \alpha^2 x$$
$$= ((1-\alpha)^2 + \alpha^2)(((1-\alpha)^2 + \alpha^2) z_{k-2} + \alpha^2 x) + \alpha^2 x = \dots$$
$$\dots = ((1-\alpha)^2 + \alpha^2)^i z_{k-i} + \alpha^2 x \sum_{j=0}^{i-1} ((1-\alpha)^2 + \alpha^2)^j$$

When $i = k$, $z_{k-i} = z_0 = 0$, and so,

$$z_k = \alpha^2 x \sum_{j=0}^{k-1} ((1-\alpha)^2 + \alpha^2)^j$$

For large enough k,

$$z_k = \frac{\alpha^2 x}{1 - ((1-\alpha)^2 + \alpha^2)} = \frac{\alpha x}{2(1-\alpha)}$$

Comparing $y_k = \frac{\alpha n x}{2-\alpha}$ with $z_k = \frac{\alpha x}{2(1-\alpha)}$, we see that for relatively small α and large n,

$$y_k \approx n z_k$$