Human in the loop RL

Sandip Patel

28 July 2020

1 Variance of the Return

Variance of any variable X can be given using the following expression:

$$V[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{1}$$

We already know the learning equation for expectation of the returns in the form of value functions and hence we have learning equation for $\mathbb{E}[x]$. In this document I am presenting the work to calculate $\mathbb{E}[X^2]$(second moment) term and finally calculate the Variance of returns in the form of Learning equations (or update rule).

2 Expectation of second moment

Let the variable x be the return by taking an action $\boldsymbol{a}$ ϵA from state $\boldsymbol{s}$ ϵS and then following the policy. Hence,

$$\mathbb{E}[x] = Q(s,a) = J(s,a) \tag{2}$$

$$Q(s,a) = r(s,a) + \gamma * \sum P(s'|s,a)Q(s',a') \tag{3}$$

our well known action value function.
Let the expectation of second moment be represented by $M(s,a)$. So,

$$\mathbb{E}[x^2] = M(s,a) \tag{4}$$

$$\begin{aligned} M(s,a) &= \mathbb{E}[returns^2|s,a] \\ &= \mathbb{E}[\ (\sum \gamma^k * r(s,a))^2 \mid s\epsilon S, a\epsilon A\ ] \\ &= \mathbb{E}[\ ((r(s_o,a_o) + \sum \gamma^k * r(s',a'))^2 \mid s_o, s'\epsilon S, a_o, a'\epsilon A\ ] \\ &= r(s_o,a_o)^2 + 2\gamma r(s_o,a_0)\sum P(s'|s_o,a_o)Q(s',a') + \gamma^2 \sum P(s'|s_o,a_o)M(s',a') \end{aligned} \tag{5}$$

Using Equation (4) one can easily find expectation of the square of the *returns*.

Update rule for *Q(s,a)* from equation (3) is well known and can be given using the following equation.

The final eqns here have the max over 'a', which I think should also be present in eqns 3 and 5

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha_q[r(s,a) + \gamma * argmax_a(Q_{t-1}(s',a')) - Q_{t-1}(s,a)] \tag{6}$$

I believe argmax should be replaced by max

Similarly, we can write the update rule for equation (5) as following.

$$M_t(s,a) = M_{t-1}(s,a) + \alpha_m[r(s,a)^2 + 2\gamma r(s,a) * argmax_a(Q_{t-1}(s',a')) + \gamma^2 M_{t-1}(s',a') - M_{t-1}(s,a)] \tag{7}$$

M_t-1 should also have an max around it.

3 Variance Update

Once we have the new estimate of Q(s,a) and M(s,a) at time step t using (6) and (7) we can use (1) to update the variance.
Since,

This term has to be squared

$$V_t(s,a) = M_t(s,a) - Q_t(s,a) \tag{8}$$

We can update the Variance using the simple Temporal difference method. Which come out to be

$$V_t(s,a) = V_{t-1}(s,a) + \alpha_v[M_t(s,a) - Q_t^2(s,a) - V_{t-1}(s,a)] \tag{9}$$