Chapter 7

Exercise 7.1:

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

N-step error:

$$G_{t:t+n} - v(S_t) = R_{t+1} + \gamma * G_{t+1:t+n} - v(S_t) + \gamma * v(S_{t+1}) - \gamma * v(S_{t+1})$$
$$= \delta_t + \gamma * [G_{t+1:t+n} - v(S_{t+1})]$$
$$= \delta_t + \gamma * \delta_{t+1} + \gamma^2 * [G_{t+2:t+n} - v(S_{t+2})]$$
$$= \delta_t + \gamma * \delta_{t+1} + \gamma^2 * \delta_{t+2} + \gamma^3 * \delta_{t+3} + ... + \gamma^{n-1} * [G_{t+n-1:t+n} - v(S_{t+n-1})]$$
$$= \delta_t + \gamma * \delta_{t+1} + \gamma^2 * \delta_{t+2} + \gamma^3 * \delta_{t+3} + ... + \gamma^{n-1} * [R_{t+n} + \gamma * v(S_{t+n}) - v(S_{t+n-1})]$$
$$= \delta_t + \gamma * \delta_{t+1} + \gamma^2 * \delta_{t+2} + \gamma^3 * \delta_{t+3} + ... + \gamma^{n-1} * \delta_{t+n-1}$$
$$= \sum_{k=t}^{t+n-1} \gamma^{k-t} * \delta_k$$

Therefore, N-Step error can be written as sum of TD errors.

Exercise 7.2 (Programming):

When state values do change every time step, the TD(0) error is given by :

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

While the N-Step TD update is:

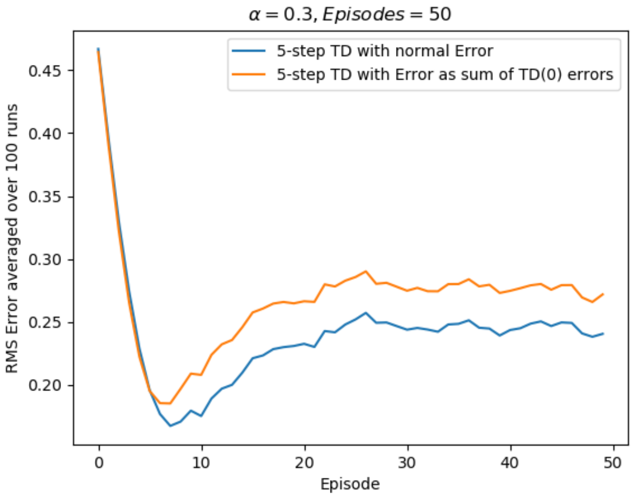
$v_{t+n}(S_t) = v_{t+n-1}(S_t) + \alpha * [G_{t:t+n} - v_{t+n-1}(S_t)]$

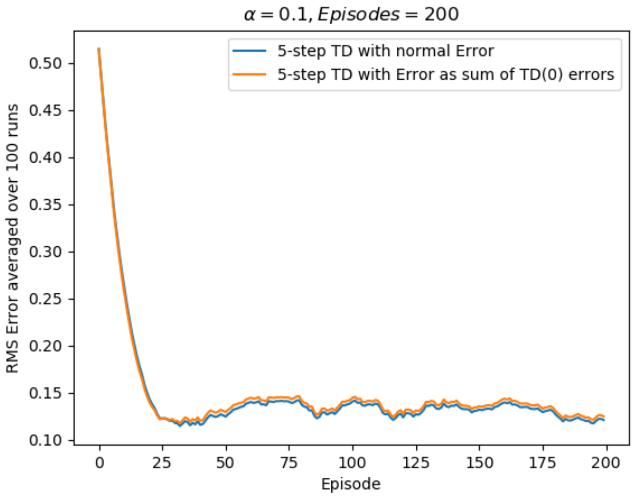
Notice that the TD error δ_t defined by the value available at time t. However, the update for state S_t at time t uses values from time t+n-1.

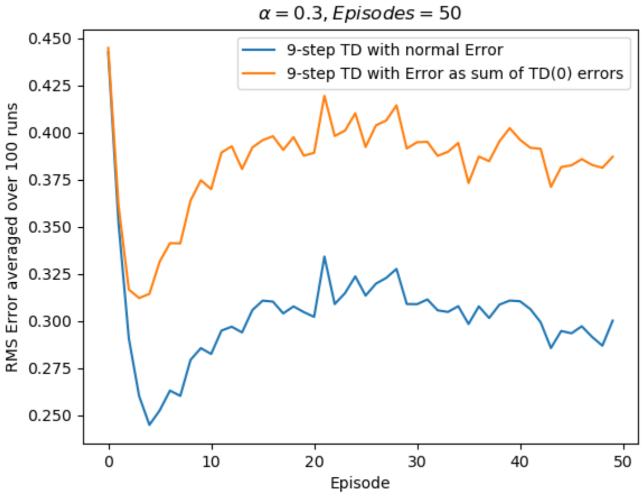
This is because we need the information of the next n steps first. Keeping this in mind, I implemented 2 different N-Step TD algorithms on a random walk with 19 states. In one algorithm I used summation of TD(0) errors to calculate the N-Step Error while the 2nd algorithm is the normal N-Step TD algorithm.

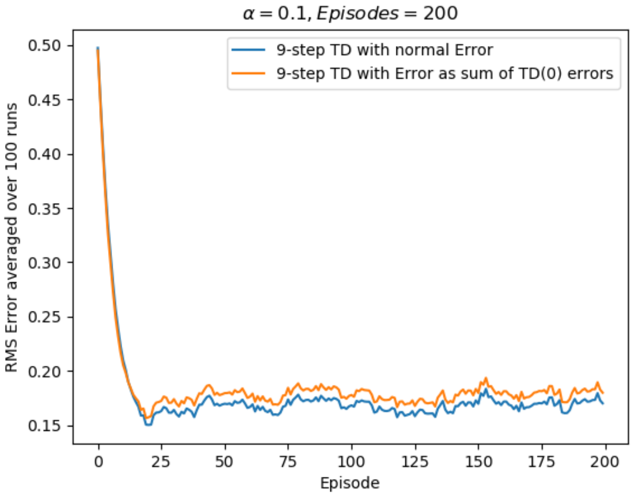
Code: Check [Github]

Results:









The two figures on the top are for the 5-Step TD algorithm while the 2 below are for the 9-Step TD algorithm.
According to the above results, the method which uses summation of TD errors is worse in comparison to the normal TD method. Further, the higher the value of N, the higher is the difference between the final error values. For the same value of N, a low value of alpha gives less difference between the final errors. This makes sense because higher the N, more is the difference between the time steps used by both methods to calculate the error. Also, lower the value of alpha, lower is the difference between the two methods, as mentioned in the book.

Exercise 7.3:

With a small number of states in the random walk, a random trajectory generated will also be small. If the trajectory length is less than ‘n’ and we are doing a n-Step TD update, it will not be possible to compare higher values of n. I think that a smaller value of n will perform better on smaller number of states and as the number of states increases, the value of best ‘n’ will also increase.

Exercise 7.4:

$$
\begin{aligned}
&G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{+n}) \\
&= R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - \gamma Q_t(S_{t+1}, A_{t+1}) + Q_{t-1}(S_t, A_t) - Q_{t-1}(S_t, A_t) \\
&+ \gamma R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) + \gamma Q_t(S_{t+1}, A_{t+1}) - \gamma Q_t(S_{t+1}, A_{t+1}) \\
&+ \gamma^2 R_{t+3} + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma^2 Q_{t+1}(S_{t+1}, A_{t+1}) \\
&+ \gamma^3 R_{t+4} + \gamma^4 Q_{t+3}(S_{t+4}, A_{t+4}) - \gamma^4 Q_{t+3}(S_{t+4}, A_{t+4}) + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) \\
&. \\
&. \\
&+ \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) - \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) + \gamma^{n-1} Q_{t+n-2}(S_{t+n-1}, A_{t+n-1}) - \gamma^{n-1} Q_{t+n-2}(S_{t+n-1}, A_{t+n-1}) \\
&+ \gamma^n Q_{t+n-1}(S_{t+n}, A_{+n})
\end{aligned}
$$

Cancel the second Q term in each row with the 3rd Q term in the next row.
Cancel the Q term in the last row with the 2nd Q term in the second last row
We get:

$$
\begin{aligned}
&G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + ...+ \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{+n}) \\
&= R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) + Q_{t-1}(S_t, A_t) - Q_{t-1}(S_t, A_t) \\
&+ \gamma R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1}) \\
&+ \gamma^2 R_{t+3} + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^2 Q_{t+1}(S_{t+1}, A_{t+1}) \\
&+ \gamma^3 R_{t+4} + \gamma^4 Q_{t+3}(S_{t+4}, A_{t+4}) - \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) \\
&. \\
&. \\
&+ \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) - \gamma^{n-1} Q_{t+n-2}(S_{t+n-1}, A_{t+n-1})
\end{aligned}
$$

Therefore,

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{min(t+n,T)-1} \gamma^{k-t}[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)$$

Exercise 7.5:

In the pseudocode given on page 149 the main update segment should be changed to:

```
If $\tau >= 0$
    $G \leftarrow 0$
    Loop $j = min(\tau + n, T) - 1$ to $j = \tau$ (reverse loop)
        $\rho \leftarrow \frac{\pi(A_j|S_j)}{b(A_j|S_j)}$
        $G \leftarrow \rho * [R_{j+1} + \gamma * G] + (1 - \rho) * V(S_j)$
        If $\tau + n < T$
            $G \leftarrow G + \gamma^n * V(S_{\tau+n})$
    $v(S_\tau) \leftarrow v(S_\tau) + \alpha * [G - v(S_\tau)]$
```

<u>Exercise 7.6:</u>

Off-Policy return with control variates is:

$G_{t:h} = R_{t+1} + \gamma[\rho_{t+1} G_{t+1:h} + \overline{v}_{h-1}(S_t) - \rho_{t+1} q(S_{t+1}, A_{t+1})]$

Expectation of this return:

$$= E[R_{t+1} + \gamma[\rho_{t+1} G_{t+1:h} + \overline{v}_{h-1}(S_t) - \rho_{t+1} q(S_{t+1}, A_{t+1})]$$
$$= E[R_{t+1}] + E[\gamma[\rho_{t+1} G_{t+1:h} + \overline{v}_{h-1}(S_t) - \rho_{t+1} q(S_{t+1}, A_{t+1})]]$$

Expected value of ρ_t is 1 and this term is independent of everything.

$$= E[R_{t+1}] + E[\gamma G_{t+1:h}] + E[\gamma \overline{v}_{h-1}(S_t)] - E[\gamma q(S_{t+1}, A_{t+1})]$$
$$= E[R_{t+1}] + E[\gamma G_{t+1:h}] + \gamma \overline{v}_{h-1}(S_t) - E[\gamma q(S_{t+1}, A_{t+1})]$$

Since $E[\gamma q(S_{t+1}, A_{t+1})] = \overline{v}(S_t)$

$$= E[R_{t+1}] + E[\gamma G_{t+1:h}]$$
$$= E[R_{t+1}] + \gamma E[R_{t+2}] + E[\gamma^2 G_{t+2:h}]$$
$$= E[R_{t+1}] + \gamma E[R_{t+2}] + \gamma^2 E[R_{t+3}] + \ldots + \gamma^{h-t-1} E[R_h] + E[\gamma^{h-t} G_{h:h}]$$
$$= R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \ldots + \gamma^{h-t-1} R_h + E(q(S_h, A_h))$$

$= E[G^{1}{}_{t:h}]$ which is the same as the return without control variate

Exercise 7.7:

```
If $\tau >= 0$
    $h \leftarrow \tau + n$
    If $h < T$
        $G \leftarrow q(S_h, A_h)$
    Else
        $G \leftarrow R_T$

    Loop $j = min(h, T - 1) - 1\ to\ j = \tau\ (reverse\ loop)$
        $\rho \leftarrow \frac{\pi(A_{j+1}|S_{j+1})}{b(A_{j+1}|S_{j+1})}$
        $G \leftarrow R_{j+1} + \gamma\rho[G - q(S_{j+1}, A_{j+1})] + \gamma\bar{v}(S_{j+1})$

    $q(S_\tau, A_\tau) \leftarrow q(S_\tau, A_\tau) + \alpha * [G - q(S_\tau, A_\tau)]$
```

Exercise 7.8:

General Off-Policy return using control variates:

$$
\begin{aligned}
G_{t:h} &= \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1 - \rho_t)v(S_t) \\
&= \rho_t(R_{t+1} + \gamma v(S_{t+1}) - v(S_t) - \gamma v(S_{t+1}) + v(S_t) + \gamma G_{t+1:h}) + (1 - \rho_t)v(S_t) \\
&= \rho_t(\delta_t - \gamma v(S_{t+1}) + v(S_t) + \gamma G_{t+1:h}) + (1 - \rho_t)v(S_t) \\
&= \rho_t(\delta_t - \gamma v(S_{t+1}) + \gamma G_{t+1:h}) + v(S_t)
\end{aligned}
$$

Let $k_t = \rho_t(\delta_t - \gamma v(S_{t+1}) + \gamma G_{t+1:h})$

$$
\begin{aligned}
G_{t:h} &= k_t + v(S_t)\ \textbf{Eq(1)} \\
&= \rho_t(\delta_t - \gamma v(S_{t+1}) + \gamma[\rho_{t+1}(R_{t+2} + \gamma G_{t+2:h}) + (1 - \rho_{t+1})v(S_{t+1})]) + v(S_t) \\
&= \rho_t(\delta_t - \gamma v(S_{t+1}) + \gamma[\rho_{t+1}(\delta_{t+1} - \gamma v(S_{t+2}) + \gamma G_{t+2:h}) + v(S_{t+1})]) + v(S_t) \\
&= \rho_t(\delta_t + \gamma[\rho_{t+1}(\delta_{t+1} - \gamma v(S_{t+2}) + \gamma G_{t+2:h})]) + v(S_t)
\end{aligned}
$$

$$= \rho_t(\delta_t + \gamma k_{t+1}) + v(S_t)$$

$G_{t:h} = \rho_t(\delta_t + \gamma k_{t+1}) + v(S_t)$ **Eq(2)**

From Eq(1) and Eq(2) we get:

$$k_t = \rho_t(\delta_t + \gamma * k_{t+1})$$

$$G_{t:h} = k_t + v(S_t)$$
$$= \rho_t(\delta_t + \gamma * k_{t+1}) + v(S_t)$$
$$= \rho_t(\delta_t + \gamma * [\rho_{t+1}(\delta_{t+1} + \gamma * k_{t+2})]) + v(S_t)$$

$$k_{h-1} = \rho_{h-1}(\delta_{h-1} - \gamma v(S_h) + \gamma G_{h:h})$$
$$k_{h-1} = \rho_{h-1}(\delta_{h-1} - \gamma v(S_h) + \gamma v(S_h))$$
$$k_{h-1} = \rho_{h-1}\delta_{h-1}$$

$$G_{t:h} = \rho_t(\delta_t + \gamma * [\rho_{t+1}(\delta_{t+1} + \gamma * k_{t+2})]) + v(S_t)$$
$$= \sum_{i=t}^{h-1}[\gamma^{i-t} * \prod_{j=t}^{j=i}\rho_j * \delta_i] + v(S_t)$$

The Error which is $G_{t:h} - v(S_t)$

Is thus given by:

$$G_{t:h} - v(S_t) = \sum_{i=t}^{h-1}[\gamma^{i-t} * \prod_{j=t}^{j=i}\rho_j * \delta_i]$$

Exercise 7.9:

Action Version of Off-Policy return:

$$G_{t:h} = R_{t+1} + \gamma[\rho_{t+1} G_{t+1:h} + \overline{v}\,(S_{t+1}) - \rho_{t+1} q(S_{t+1}, A_{t+1})]$$

Expected SARSA TD Error:

$$\delta_t = R_{t+1} + \gamma E[v(S_{t+1})] - q(S_t, A_t)$$
$$\delta_t = R_{t+1} + \gamma \overline{v}(S_{t+1}) - q(S_t, A_t)$$

$$\begin{aligned} G_{t:h} = {} & R_{t+1} + \gamma \overline{v}(S_{t+1}) - q(S_t, A_t) \\ & - \gamma \overline{v}(S_{t+1}) + q(S_t, A_t) + \gamma[\rho_{t+1} G_{t+1:h} + \overline{v}\,(S_{t+1}) - \rho_{t+1} q(S_{t+1}, A_{t+1})] \end{aligned}$$

$= \delta_t + q(S_t, A_t) + \gamma[\rho_{t+1} G_{t+1:h} - \rho_{t+1} q(S_{t+1}, A_{t+1})]$ **Eq(1)**

Let $k_t = \rho_{t+1}[G_{t+1:h} - q(S_{t+1}, A_{t+1})]$

$$G_{t:h} = \delta_t + q(S_t, A_t) + \gamma k_t$$

$$G_{t+1:h} = \delta_{t+1} + q(S_{t+1}, A_{t+1}) + \gamma[\rho_{t+2} G_{t+2:h} - \rho_{t+2} q(S_{t+2}, A_{t+2})]$$
$$G_{t+1:h} = \delta_{t+1} + q(S_{t+1}, A_{t+1}) + \gamma k_{t+1}$$

Substituting in **Eq(1):**

$$G_{t:h} = \delta_t + q(S_t, A_t) + \gamma[\rho_{t+1} \delta_{t+1} + \gamma \rho_{t+1} k_{t+1}]$$
$$\begin{aligned} G_{t:h} = {} & \delta_t + q(S_t, A_t) \\ & + \gamma[\rho_{t+1} \delta_{t+1} + \gamma \rho_{t+1} \{\rho_{t+2} G_{t+2:h} - \rho_{t+2} q(S_{t+2}, A_{t+2})\}] \end{aligned}$$

$$\begin{aligned} G_{t:h} = {} & \delta_t + q(S_t, A_t) \\ & + \gamma[\rho_{t+1} \delta_{t+1} + \gamma \rho_{t+1} \{\rho_{t+2}[\delta_{t+2} + q(S_{t+2}, A_{t+2}) + \gamma k_{t+2}] - \rho_{t+2} q(S_{t+2}, A_{t+2})\}] \end{aligned}$$

$$G_{t:h} = \delta_t + q(S_t, A_t) + \gamma[\rho_{t+1}\delta_{t+1} + \gamma\rho_{t+1}\{\rho_{t+2}[\delta_{t+2} + \gamma k_{t+2}]]$$
$$G_{t:h} = \delta_t + q(S_t, A_t) + \gamma[\rho_{t+1}\delta_{t+1} + \gamma\rho_{t+1}\rho_{t+1}[\delta_{t+2} + \ldots + (\delta_{h-1} + \gamma k_{h-1})]]$$

$$k_{h-1} = \rho_h G_{h:h} - \rho_h q(S_h, A_h)$$
$$k_{h-1} = \rho_h q(S_h, A_h) - \rho_h q(S_h, A_h)$$
$$k_{h-1} = 0$$

Therefore,

$$G_{t:h} = q(S_t, A_t) + \sum_{i=t}^{h-1}[\gamma^{i-t}(\prod_{j=t+1}^{j=i} \rho_j)\delta_i]$$

TD Error:

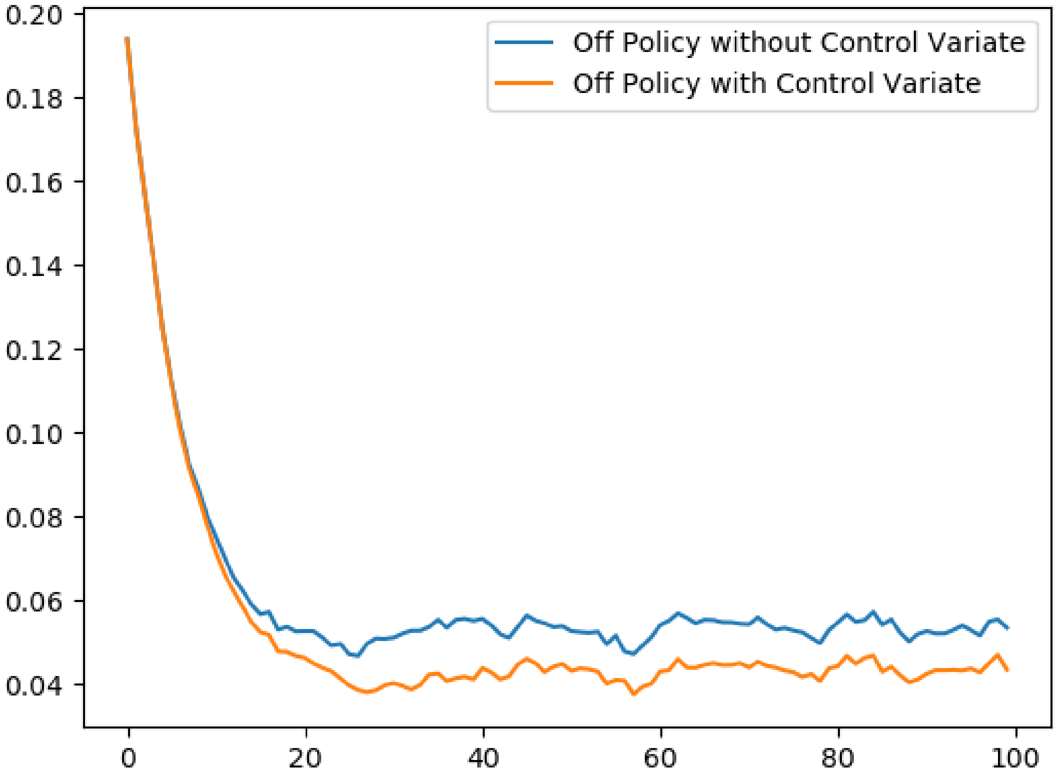
$$G_{t:h} - q(S_t, A_t) = \sum_{i=t}^{h-1}[\gamma^{i-t}(\prod_{j=t+1}^{j=i} \rho_j)\delta_i]$$

Exercise 7.10 (Programming):

I ran both the algorithms on a random walk with a target policy which selects left or right action with a random probability and a behaviour policy which chooses between left and right randomly.

Code: Check Github

Result:



Exercise 7.11:

Tree Backup return:

$$G_{t:t+n} = R_{t+1} + \sum_{a \neq A_{t+1}} \pi(a|S_{t+1})q(S_{t+1}, a) + \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n}$$

Expected SARSA TD Error:

$$\delta_t = R_{t+1} + \gamma E[v(S_{t+1})] - q(S_t, A_t)$$

$$\delta_t = R_{t+1} + \gamma\bar{v}(S_{t+1}) - q(S_t, A_t)$$

$$G_{t:t+n} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1})q(S_{t+1}, a) + \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n}$$

$$= \delta_t - \gamma\bar{v}(S_{t+1}) + q(S_t, A_t) + \sum_{a \neq A_{t+1}} \pi(a|S_{t+1})q(+ \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n}$$

We can write

$$\sum_{a \neq A_{t+1}} \pi(a|S_{t+1})q(S_{t+1}, a) \quad \text{as} \quad \bar{v}(S_{t+1}) - \pi(A_{t+1}|S_{t+1})q(S_{t+1}, A_{t+1})$$

$$\begin{aligned} G_{t:t+n} = {} & \delta_t - \gamma\bar{v}(S_{t+1}) + q(S_t, A_t) \\ & + \gamma\bar{v}(S_{t+1}) - \gamma\pi(A_{t+1}|S_{t+1})q(S_{t+1}, A_{t+1}) + \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n} \end{aligned}$$

$= \delta_t + q(S_t, A_t) - \gamma\pi(A_{t+1}|S_{t+1})q(S_{t+1}, A_{t+1}) + \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n}$ **Eq(1)**

Let $k_t = \delta_t - \gamma\pi(A_{t+1}|S_{t+1})q(S_{t+1}, A_{t+1}) + \gamma\pi(A_{t+1}|S_{t+1})G_{t+1:t+n}$

$$G_{t+1:t+n} = \delta_{t+1} + q(S_{t+1}, A_{t+1}) - \gamma\pi(A_{t+2}|S_{t+2})q(S_{t+2}, A_{t+2}) + \gamma\pi(A_{t+2}|S_{t+2})G_{t+2:t+n}$$

Substituting in **Eq(1):**

$$\begin{aligned} & = \delta_t + q(S_t, A_t) - \gamma\pi(A_{t+1}|S_{t+1})q(S_{t+1}, A_{t+1}) \\ & + \gamma\pi(A_{t+1}|S_{t+1})[\delta_{t+1} + q(S_{t+1}, A_{t+1}) - \gamma\pi(A_{t+2}|S_{t+2})q(S_{t+2}, A_{t+2}) + \gamma\pi(A_{t+2}|S_{t+2})G_{t+2:t+n}] \end{aligned}$$

$$\begin{aligned} & = \delta_t + q(S_t, A_t) \\ & + \gamma\pi(A_{t+1}|S_{t+1})[\delta_{t+1} - \gamma\pi(A_{t+2}|S_{t+2})q(S_{t+2}, A_{t+2}) + \gamma\pi(A_{t+2}|S_{t+2})G_{t+2:t+n}] \\ & = \delta_t + q(S_t, A_t) + \gamma\pi(A_{t+1}|S_{t+1})k_{t+1} \end{aligned}$$

$$\begin{aligned} & G_{t:t+n} = q(S_t, A_t) + k_t = \delta_t + q(S_t, A_t) + \gamma\pi(A_{t+1}|S_{t+1})k_{t+1} \\ & \Rightarrow k_t = \delta_t + \gamma\pi(A_{t+1}|S_{t+1})k_{t+1} \end{aligned}$$

$$G_{t:t+n} = q(S_t, A_t) + k_t$$
$$G_{t:t+n} = q(S_t, A_t) + \delta_t + \gamma\pi(A_{t+1}|S_{t+1})k_{t+1}$$
$$G_{t:t+n} = q(S_t, A_t) + \delta_t + \gamma\pi(A_{t+1}|S_{t+1})[\delta_{t+1} + \gamma\pi(A_{t+2}|S_{t+2})k_{t+2}]$$

$$k_{t+n-1} = \delta_{t+n-1} - \gamma\pi(A_{t+n}|S_{t+n})q(S_{t+n}, A_{t+n}) + \gamma\pi(A_{t+n}|S_{t+n})G_{t+n:t+n}$$
$$k_{t+n-1} = \delta_{t+n-1} - \gamma\pi(A_{t+n}|S_{t+n})q(S_{t+n}, A_{t+n}) + \gamma\pi(A_{t+n}|S_{t+n})q(S_{t+n}, A_{t+n})$$
$$k_{t+n-1} = \delta_{t+n-1}$$

$$G_{t:t+n} =$$
$$q(S_t, A_t) + \delta_t + \gamma\pi(A_{t+1}|S_{t+1})[\delta_{t+1} + ... + \gamma\pi(A_{t+n-1}|S_{t+n-1})k_{t+n-1}]$$
$$= q(S_t, A_t) + \delta_t + \gamma\pi(A_{t+1}|S_{t+1})[\delta_{t+1} + ... + \gamma\pi(A_{t+n-1}|S_{t+n-1})\delta_{t+n-1}]$$

$$G_{t:t+n} = \sum_{i=t}^{t+n-1} [\gamma^{i-t} * [\prod_{j=t+1}^{i} \pi(A_{t+1}|S_{t+1})] * \delta_i] + q(S_t, A_t)$$

The Tree Backup Error is thus:

$$G_{t:t+n} - q(S_t, A_t) = \sum_{i=t}^{t+n-1} [\gamma^{i-t} * [\prod_{j=t+1}^{i} \pi(A_{t+1}|S_{t+1})] * \delta_i]$$