1 a) Weighted Average

$$V_n = \frac{\sum_{k=1}^{n-1} w_k G_k}{\sum_{k=1}^{n-1} w_k}$$

$$V_{n+1} = \frac{\sum_{k=0}^{n} w_k G_k}{\sum_{k=0}^{n} w_k}$$

$$= \frac{1}{\sum_{k=0}^{n} w_k} \times \quad \left(= \frac{1}{C_n}\right)$$

$$= \frac{1}{C_n}\left(w_n G_n + \sum_{k=0}^{n-1} w_k G_k\right)$$

where $\sum_{k=0}^{n-1} w_k G_k = \left(\sum_{k=0}^{n-1} w_k\right) V_n$

$$= \sum_{k=0}^{n} w_k - w_n$$

$$\therefore V_{n+1} = V_n + \frac{w_n}{\sum_{k=0}^{n} w_k}\left(G_n - V_n\right)$$

Cn can covered in the same method

$$\Rightarrow C_n = \sum_{k=0}^{n} w_k = w_n - \sum_{k=0}^{n-1} w_k$$

$$\therefore V_{n+1} = V_n + \frac{w_n}{C_n}\left[G_n - V_n\right], \; n \geq 1$$

1b)
```
Initialize, for all $s \in S$, $a \in A(S)$
        $Q(s,a) \in R$ (arbitrarily)
        $C(s,a) \leftarrow 0$
        $\pi(s) \leftarrow \arg\max_a Q(s,a)$ (with ties brocken consistently)
    Loop forever (for each episode):
        $b \leftarrow$ any soft policy
        Generate an episode using $b$: $S_0, A_0, R_1 \ldots S_{T-1}, A_{T-1}, R_T$
        $G \leftarrow 0$
        $W \leftarrow 1$
        Loop for each step of episode $t = T-1, T-2, \ldots 0$
            $G \leftarrow G + R_{t+1}$
$\omega = 0$ after $A_t \neq \pi(S_t)$ {
            $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
            $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
            $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties brocken constantly)
            If $A_t \neq \pi(S_t)$ then exit inner loop
$=0$ for all $t$ after $A_t \neq \pi(S_t)$ {
            $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

The use of $1/b(A_t|S_t)$ instead of $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ in the update of W is correct because W is itself an accumulation of the importance-sampling ratios from the beginning of the episode to the time t, and adjusts the update at each time step. The policy π comes into play by determining which actions ~~not supported~~ contributed to the update if $\pi(A_t|S_t)$ is zero. then the update would be zero, effectively filtering out the updates from actions not supported by the target policy π, even though the data is generated by following behavior policy b.

3a) The agent must have moved to left terminal from state A receiving no reward, and since the terminal state has a value of 0:

$$V(A) \longleftarrow 0.5 + 0.1\,[0 + 1(0) - 0.5]$$

$$= 0.5 - 0.05 = 0.45 \,/\!/$$

All the other states were updated as well bating ~~having~~ the same reward and same initialization the error update in TD was very small

Ex:-

$$v(B) = 0.5$$

$$V(c) \longleftarrow 0.5 + 0.1\,[0 + 1\;0.5 - 0.5]$$

$$= 0.5$$