1. (a)

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated
Initialize:
  $V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$
  $Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$ $N(s) \leftarrow 0$
Loop forever (for each episode):
  Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$ $N(S_t) \leftarrow N(S_t) + 1$
    Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:
      ~~Append G to $Returns(S_t)$~~
      ~~$V(S_t) \leftarrow$ average($Returns(S_t)$)~~

$$V(S_t) \longleftarrow V(S_t) + \frac{1}{N(S_t)}\left[G_t - V(S_t)\right]$$

(b)

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
  $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
  $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
  $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$ $N(s,a) \leftarrow 0$
Loop forever (for each episode):
  Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
  Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      ~~Append G to $Returns(S_t, A_t)$~~ $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      ~~$Q(S_t, A_t) \leftarrow$ average($Returns(S_t, A_t)$)~~
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

$$Q(S_t, A_t) \longleftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}\left[G - Q(S_t, A_t)\right]$$

2. (a) There is no difference between using every-visit MC and first-visit MC because all states in a episode will only occur once. So the return for each occurrence should be the same as it for the first occurrence.

(b)

$G_0 = 10 \quad G_1 = 9 \quad \cdots \quad G_9 = 1 \quad G_{10} = 0$

First-visit: $V_\pi = \frac{G_0}{1} = 10$ Every-visit: $V_\pi = \frac{0+1+\cdots+9+10}{10} = 5.5$

4.

(c) $\varepsilon = 0$ means policy doesn't explore any more. As we can see, the return isn't increased during the episodes. It cannot be optimal. Exploring starts make the probabilities of all states higher then 0. That makes the policy attempt to take others action, instead of the first positive result.

5. (1)

$$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k} = \frac{W_n \cdot G_n + \sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k}$$

$$= \frac{W_n G_n}{\sum_{k=1}^{n} W_k} + \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k}$$

$$= \frac{W_n G_n}{\sum_{k=1}^{n} W_k} + \frac{\sum_{k=1}^{n-1} W_k}{\sum_{k=1}^{n-1} W_k} \cdot \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k}$$

$$= \frac{W_n G_n}{\sum_{k=1}^{n} W_k} + V_n \cdot \frac{\sum_{k=1}^{n-1} W_k + (W_n - W_n)}{\sum_{k=1}^{n} W_k}$$

$$= \frac{W_n G_n}{\sum_{k=1}^{n} W_k} + V_n - \frac{W_n}{\sum_{k=1}^{n} W_k} = V_n + \frac{W_n (G_n - V_n)}{\sum_{k=1}^{n} W_k}$$

Set $C_n = \sum_{k=1}^{n} W_k = C_{n-1} + W_n$

rewrite the equation $= V_n + \dfrac{W_n (G_n - V_n)}{C_n}$

(2) Because policy π is greedy. $A_t = \pi(S_t)$. Then $\pi(A_t | S_t) = 1$

Therefore, it involve $\dfrac{1}{b(A_t | S_t)}$