Question 1 (Exercise 5.4)

$$\cancel{Q_t(q,a) = R_{t+1} + \gamma R_{t+2}}$$

$$Q_t(q,a) = (G_1 + G_2 - - + G_t)/t$$

$$Q_{t+1}(q,a) = (G_1 + \ldots G_t + \cancel{G_{t+1}}) / t \cancel{+1}$$

$$= \frac{G_1 + \ldots + G_{t-1}}{t \cancel{+1}} + \frac{G_{t \cancel{+1}}}{t \cancel{+1}}$$

$$= \left(\frac{1}{t \cancel{+1}}\right)\left[t-1\left(\frac{G_1 + \ldots G_t}{t-1}\right) + \frac{G_{t+1}}{\cancel{t+1}}\right]$$

$$= \left(\frac{1}{t \cancel{+1}}\right)\left[(t-1) Q_t(q,a) + \frac{G_{t+1}}{\cancel{t+1}}\right]$$

$$t Q_t(q,a) - Q_t(q,a)$$

$$Q_{t+1}(q,a) = \left(\frac{1}{t \cancel{+1}}\right)^2 \left[\cancel{t(t-1) Q(q,a)} + G_{t+1}\right]$$

$$= Q_t(q,a) - \frac{(G_t - Q_t(q,a))}{t} \qquad \Big| \; G_t \Leftrightarrow R_t$$

G_i → It represent return when state-action pair is visit first in and then discounted average return from there.

$Q_T(q,a)$ → Action-state value after T'th update of state-action pair (q,a)

~~t → It is t~~

t → It is number of times current state-action pair has been updated. Excluding current update

Initialize
$\pi(s) \in A(s)$
$Q(s,a) \in \mathbb{R}$
Returns $(s,a) \leftarrow$ empty list, for all $s \in S$, $a \in A(s)$
$n(s,a) = 1$ $\quad s \in S, a \in A(s)$

~~Loop~~
Loop forever
Choose $S_0 \in S$ $A_0 \in A(S_0)$ randomly such that all pair 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, $t = T-1, T-2, \ldots 0$
$G \leftarrow \gamma G + R_{t+1}$

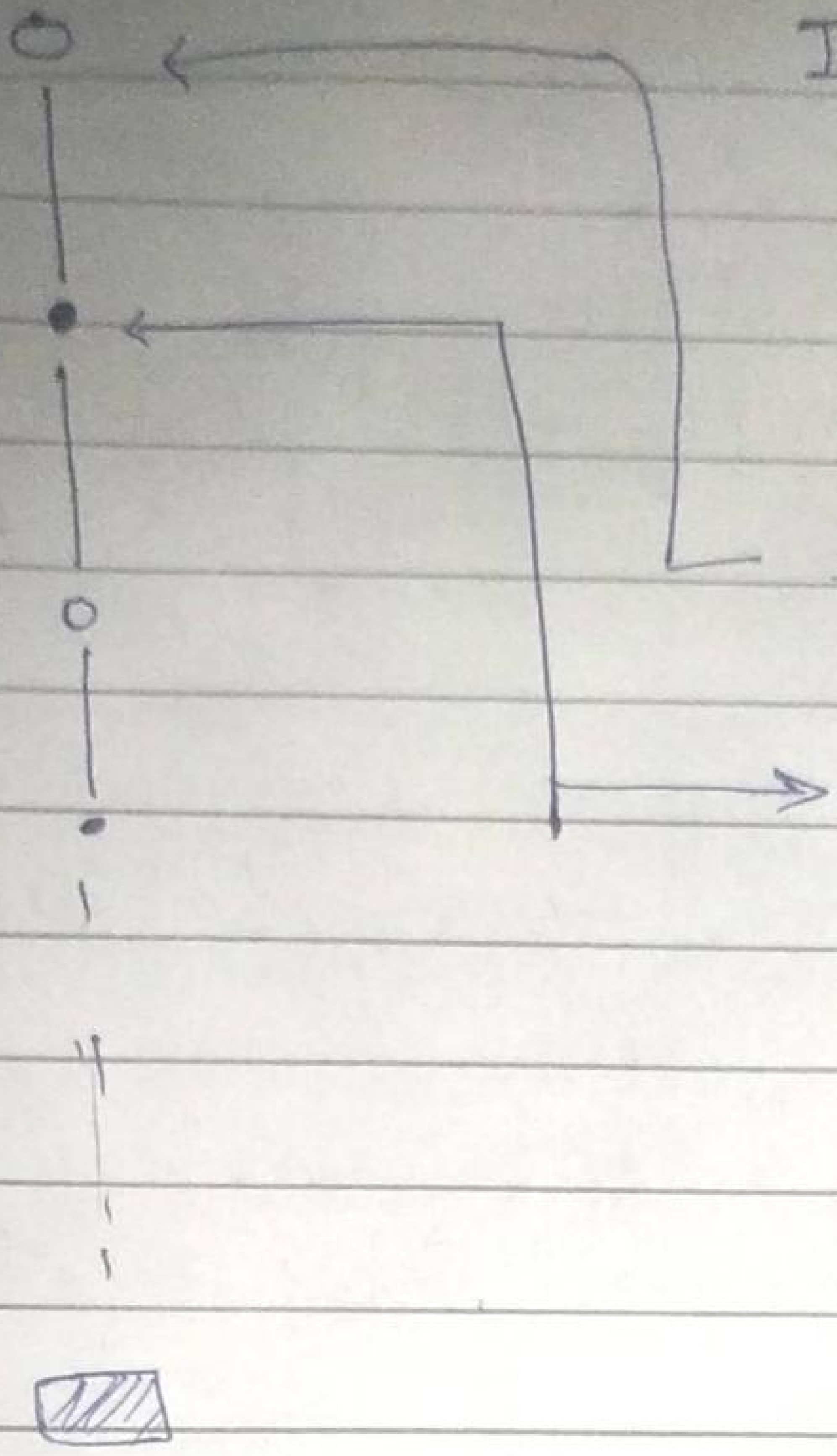
Unless the pair S_t, A_t appear in $S_0, A_0, S_1 \ldots S_{t-1}, A_{t-1}$
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{n(S_t,a)} \left[G - Q(S_t, A_t) \right]$$
$$n(s_t, a_t) = n(s_t, a_t) + 1$$

It is single visit & Incremental approach

Question 2.
Exersice 5.3:

If $q_\pi(s', a')$ is show by left diagram

Then



Question 5 (Exercise 6.2):

Any task which is completely markov in nature.
Like:-

Moves in chess. Given we are in certain state in chess and have experiance then we can predict our chances of wining without awaiting to game.
End

Question 3 (Exercise 5.6)

Single visit MC :-

Assume at time t we are in states S_t and taking action A_t. Now, probability of subsequent trajectory

trajectory $\{(S_{t+1}, A_{t+1}) \ldots (S_{T-1}, A_{T-1}), S_T\}$

$$P_r\{\,\cdot \mid S_t, A_t; \pi\} = \cancel{\prod p(S_{t+1} \mid S_t, A_t)\, p(A_{t+1} \mid S_{t+1})}$$

$$= \cancel{\prod p(S_k \mid S_{k-1}, A_k}$$

$$= \left[\prod_{k=t}^{T-1} p(S_{k+1} \mid S_k, A_k)\, \pi(A_{k+1} \mid S_{k+1})\right] p(S_T \mid S_{T-1}, A_{T-1})$$

So, Sampling Importance Ratio

$$= \prod_{k=t}^{T-1} \frac{\pi(A_{k+1} \mid S_{k+1})}{b(A_{k+1} \mid S_{k+1})}$$

Replace this Ratio in Eq of off-line state value.

$$V(s,a) = \frac{\sum_{t \in \tau(s,a)} \rho_{t:T(t)-1}\, G_t}{\sum_{t \in \tau(s,a)} \rho_{t:T(t)-1}}$$

$\tau(s,a)$ mean in pair (s,a)

~~Consider~~ $\cancel{A_{k+1}}$,

Consider $b(A_i \mid S_i)$ only after ~~pass~~ taking action A in state S.

Question 8 (Exercise 6.12) Page 5

No,
In Greedy SARSA behaviour policy is dynamic.
But, In offline q-learning, policy is static.