E1 277 Reinforcement Learning

Assignment - 2

-KAWIN M

Question 1 (a)

Sample complexity of Q-value iteration:

$$\| Q_K - Q^* \|_\infty \leq \epsilon$$

$$\Rightarrow \| Q_K - T^* Q^* \|_\infty \leq \epsilon$$ [$\because Q^*$ is a fixed point of T^* $\Rightarrow Q^* = T^* Q^*$]

$$\Rightarrow \| T^* Q_{K-1} - T^* Q^* \|_\infty \leq \epsilon$$ [Using Q-value iteration algorithm $\Rightarrow Q_K = T^* Q_{K-1}$]

Expanding $T^* Q_{K-1}$, $T^* Q^*$ using the definitions of Bellman optimal operators.

$$\Rightarrow \| r(s,a) + \gamma \int P(s'|s,a) \max_{a' \in A} Q_{K-1}(s',a') ds' - \left[r(s,a) + \gamma \int P(s'|s,a) \max_{a' \in A} Q^*(s',a') ds' \right] \|_\infty \leq \epsilon$$

Cancelling out $r(s,a)$

$$\Rightarrow \| \gamma \int P(s'|s,a) \max_{a' \in A} Q_{K-1}(s',a') ds' - \gamma \int P(s'|s,a) \max_{a' \in A} Q^*(s',a^*) ds' \|_\infty \leq \epsilon$$

$$\Rightarrow \| \gamma \left[\int P(s'|s,a) \max_{a' \in A} Q_{K-1}(s',a') ds' - \int P(s'|s,a) \max_{a' \in A} Q^*(s',a') ds' \right] \|_\infty \leq \epsilon$$

$$\Rightarrow \gamma \left\| \int P(s'|s,a) \max_{a' \in A} \left[Q_{k-1}(s',a') - Q^*(s',a') \right] ds' \right\|_\infty \le \epsilon$$

$$\Rightarrow \gamma \left\| \max_{a' \in A} \left[Q_{k-1}(s',a') - Q^*(s',a') \right] \int P(s'|s,a)\, ds' \right\|_\infty \le \epsilon$$

$$\Rightarrow \gamma \left\| \max_{a' \in A} \left[Q_{k-1}(s',a') - Q^*(s',a') \right] \cdot 1 \right\|_\infty \le \epsilon$$

using definition of max-norm

$$\Rightarrow \gamma \max_{\substack{s' \in S, \\ a' \in A}} \left[\max_{a' \in A} \left[Q_{k-1}(s',a') - Q^*(s',a') \right] \right] \le \epsilon$$

$$\Rightarrow \gamma \max_{s' \in S,\, a' \in A} \left[Q_{k-1}(s',a') - Q^*(s',a') \right] \le \epsilon$$

$$\Rightarrow \gamma \left\| Q_{k-1}(s',a') - Q^*(s',a') \right\|_\infty \le \epsilon$$

$$\Rightarrow \gamma \| Q_{k-1} - Q^* \|_\infty \le \epsilon$$

$$\Rightarrow \| Q_k - Q^* \|_\infty \le \gamma \| Q_{k-1} - Q^* \|_\infty \le \epsilon$$

Doing this iteratively

$$\le \gamma^2 \| Q_{k-2} - Q^* \|_\infty \le \epsilon$$

$$\vdots$$

$$\le \gamma^k \| Q_0 - Q^* \|_\infty \le \epsilon \longrightarrow (1)$$

We know that,

$$Q(s,a) = E[G_t \mid S_t = s, A_t = a] \le G_t^{max}$$

and $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$

$$G_t^{max} \triangleq R_{max} + \gamma R_{max} + \gamma^2 R_{max} + \cdots$$

$$= R_{max} [1 + \gamma + \gamma^2 + \cdots]$$

$$= R_{max} \cdot \frac{1}{1-\gamma}$$

[R_{max} = Maximum possible reward value]

$$G_t^{max} = \frac{R_{max}}{(1-\gamma)}$$

$$\therefore Q(s,a) \leq \frac{R_{max}}{(1-\gamma)}$$

$$\Rightarrow \|Q_0 - Q^*\|_\infty \leq \frac{R_{max}}{(1-\gamma)} \longrightarrow (2)$$

$$\Rightarrow \gamma^K \|Q_0 - Q^*\|_\infty \leq \gamma^K \frac{R_{max}}{(1-\gamma)} \leq \epsilon$$

$$\Rightarrow \gamma^K \leq \frac{\epsilon(1-\gamma)}{R_{max}}$$

Taking log on both sides

$$\log \gamma^K \leq \log\left(\frac{(1-\gamma)\epsilon}{R_{max}}\right)$$

$$K \log \gamma \leq \log\left(\frac{(1-\gamma)\epsilon}{R_{max}}\right)$$

$$K \leq \frac{\log\left(\frac{(1-\gamma)\epsilon}{R_{max}}\right)}{\log \gamma}$$

$\therefore$ The upper bound on number of iterations K required to ensure that the error $\|Q_K - Q^*\|_\infty \leq \epsilon$

is

$$k \leq \frac{\log\left(\frac{(1-\gamma)\epsilon}{R_{max}}\right)}{\log \gamma}$$

Question 1 (b)

To compute the sample complexity of the policy iteration algorithm to ensure that $\|Q^{\pi_K} - Q^*\|_\infty \leq \epsilon$

$$\|Q^{\pi_K} - Q^*\|_\infty \leq \epsilon$$

$$\Rightarrow \|Q^{\pi_K} - T^* Q^*\|_\infty \leq \epsilon$$ [$\because Q^*$ is a fixed point of T^* $\Rightarrow Q^* = T^* Q^*$]

$$\Rightarrow \|T^{\pi} Q^{\pi_{K-1}} - T^* Q^*\|_\infty \leq \epsilon$$ [Using Policy Evaluation $\Rightarrow Q^{\pi_K} = T^{\pi} Q^{\pi_{K-1}}$]

Expanding $T^{\pi} Q^{\pi_{K-1}}$ and $T^* Q^*$.

$$\Rightarrow \| r(s,a) + \gamma \int P(s'|s,a) \int \pi(a'|s') Q^{\pi_{K-1}}(s',a')\, da'\, ds'$$

$$- r(s,a) - \gamma \int P(s'|s,a) \max_{a' \in A} Q^*(s',a')\, ds' \|_\infty \leq \epsilon$$

Cancelling out $r(s,a)$

$$\Rightarrow \| \gamma \int P(s'|s,a) \left[\int \pi(a'|s') Q^{\pi_{K-1}}(s',a')\, da' - \max_{a' \in A} Q^*(s',a') \right] ds' \|_\infty \leq \epsilon$$

$$\leq \gamma \| \int P(s'|s,a) \left[\int \pi(a'|s') \max_{a' \in A} Q^{\pi_{K-1}}(s',a')\, da' - \max_{a' \in A} Q^*(s',a') \right] ds' \|_\infty \leq \epsilon$$

$$\leq \gamma \| \int P(s'|s,a) \left[\max_{a' \in A} Q^{\pi_{K-1}}(s',a') \int \pi(a'|s')\, da' - \max_{a' \in A} Q^*(s',a') \right] ds' \|_\infty \leq \epsilon$$

$$\leq \gamma \left\| \int P(s'|s,a) \left[\max_{a' \in A} \left(Q^{\pi_{k-1}}(s',a') - Q^*(s',a') \right) \right] ds' \right\|_\infty$$
$$\leq \epsilon$$

$$\leq \gamma \left\| \max_{s',a'} \left(Q^{\pi_{k-1}}(s',a') - Q^*(s',a') \right) \int P(s'|s,a) \, ds' \right\|_\infty \leq \epsilon$$

$$\leq \gamma \left\| \max_{s',a'} \left(Q^{\pi_{k-1}}(s',a') - Q^*(s',a') \right) \cdot 1 \right\|_\infty \leq \epsilon$$

$$\leq \gamma \| Q^{\pi_{k-1}} - Q^* \|_\infty \leq \epsilon$$

$$\Rightarrow \| Q^{\pi_k} - Q^* \|_\infty \leq \gamma \| Q^{\pi_{k-1}} - Q^* \|_\infty \leq \epsilon$$

Doing this iteratively, we get

$$\leq \gamma^2 \| Q^{\pi_{k-2}} - Q^* \|_\infty \leq \epsilon$$

$$\vdots$$

$$\leq \gamma^k \| Q^{\pi_0} - Q^* \|_\infty \leq \epsilon$$

Using Equation (2) from Question (1)(a)

$$\Rightarrow \| Q^{\pi_k} - Q^* \|_\infty \leq \gamma^k \frac{R_{max}}{(1-\gamma)} \leq \epsilon$$

$$\Rightarrow \gamma^k \leq \frac{\epsilon(1-\gamma)}{R_{max}}$$

Taking log on both sides

$$\Rightarrow k \log \gamma \leq \log \left(\frac{\epsilon(1-\gamma)}{R_{max}} \right)$$

$$\Rightarrow \boxed{k \leq \frac{\log \left(\frac{\epsilon(1-\gamma)}{R_{max}} \right)}{\log \gamma}}$$

Question 1 (c)

Considering all the non-stationary policies $\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \ldots$ over the state space S and action space A.

Let the order of the value functions corresponding to the non-stationary policies be

$$v^{\tilde{\pi}_1} \leq v^{\tilde{\pi}_2} \leq v^{\tilde{\pi}_3} \leq \ldots$$

It is given that

$$v^{+} = \sup_{\tilde{\pi}} v^{\tilde{\pi}}$$

$$\therefore v^{\tilde{\pi}_1} \leq v^{\tilde{\pi}_2} \leq v^{\tilde{\pi}_3} \leq \ldots \ldots \leq v^{+} \longrightarrow (1)$$

But, we also know that

$$v^{\tilde{\pi}_1} \leq v^{\tilde{\pi}_2} \leq v^{\tilde{\pi}_3} \leq \ldots \ldots \leq v^{*} \longrightarrow (2)$$

From (1) & (2)

$$v^{+} \leq v^{*} \quad \text{and} \quad v^{*} \leq v^{+}$$

$$\Rightarrow v^{+} = v^{*}$$

Hence Proved

Conclusion:

It can be concluded that even in the continuous state and action space and with non-stationary policies, repeated application of the Bellman operator with supremum value converges to the optimal value.

Question ① ⓓ

Given: $(\hat{T}^{\pi} V)(s) = \gamma^{\pi}(s) \int P(s'|s) V(s') ds'$

i) Monotonicity:

To Prove: For any $V, \bar{V} \in \mathbb{R}^n$, such that, let

$$V(i) \le \bar{V}(i) \quad \forall\ 1 \le i \le n$$

then, $(\hat{T}^{\pi}_{K} V)(i) \le (\hat{T}^{\pi}_{K} \bar{V})(i)$

Proof using Induction:

Base case: $K = 1$

$$(\hat{T}^{\pi} V)(i) = \gamma^{\pi}(i) \int P(s'|i) V(s') ds'$$

$$\le \gamma^{\pi}(i) \int P(s'|i) \bar{V}(s') ds' \quad [\because V(i) \le \bar{V}(i)]$$

$$\le (\hat{T}^{\pi} \bar{V})(i)$$

Hence Proved

Induction Hypothesis: Let us assume that for any arbitrary K,

$(\hat{T}^{\pi}_{K} V)(i) \le (\hat{T}^{\pi}_{K} \bar{V})(i)$ is true

To Prove: $(\hat{T}^{\pi}_{K+1} V)(i) \le (\hat{T}^{\pi}_{K+1} \bar{V})(i)$

Proof: $(\hat{T}^{\pi}_{K+1} V)(i) = (\hat{T}^{\pi} V^{\pi}_{K})(i)$

$$= \gamma^{\pi}(i) \int P(s'|i) V^{\pi}_{K}(s') ds'$$

$$= \gamma^{\pi}(i) \int P(s'|i) (\hat{T}^{\pi}_{K} V)(s') ds'$$

$$\leq \gamma^{\pi}(i) \int P(s'|i) (\hat{T}^{\pi}_{k} \bar{V})(s') ds' \quad [\because \text{Induction Hypothesis}]$$

$$\leq \gamma^{\pi}(i) \int P(s'|i) (\bar{V}^{\pi}_{k})(s') ds'$$

$$\leq (\hat{T}^{\pi}_{k+1} \bar{V})(i)$$

Henced Proved that the given Bellman operator is monotonic using Induction

(ii) contraction:

To Prove: $\| \hat{T}^{\pi} V - \hat{T}^{\pi} \bar{V} \|_{\infty} \leq \beta \| V - \bar{V} \|_{\infty}$,

for any $V, \bar{V} \in \mathbb{R}^n$ and a scalar $0 < \beta < 1$

Proof:

$$\hat{T}^{\pi} V - \hat{T}^{\pi} \bar{V} = \gamma^{\pi}(s) \int P(s'|s) V(s') ds' - \gamma^{\pi}(s) \int P(s'|s) \bar{V}(s') ds'$$

$$= \gamma^{\pi}(s) \int P(s'|s) [V(s') - \bar{V}(s')] ds'$$

$$|\hat{T}^{\pi} V - \hat{T}^{\pi} \bar{V}| = |\gamma^{\pi}(s) \int P(s'|s) [V(s') - \bar{V}(s')] ds'|$$

$$= \gamma^{\pi}(s) \int P(s'|s) |V(s') - \bar{V}(s')| ds'$$

Taking $\max_{i=s..n}$ on both sides

$$\max_{s=1..n} |\hat{T}^{\pi} V - \hat{T}^{\pi} \bar{V}| \leq \gamma^{\pi}(s) \int P(s'|s) \max_{s=1..n} |V(s') - \bar{V}(s')| ds'$$

$$\Rightarrow \| \hat{T}^{\pi} V - \hat{T}^{\pi} \bar{V} \|_{\infty} \leq \gamma^{\pi}(s) \max_{s=1..n} |V(s') - \bar{V}(s')| \int P(s'|s) ds'$$

$$\|\hat{T}^{\pi} v - \hat{T}^{\pi} \bar{v}\|_{\infty} \leq \gamma^{\pi}(s) \|v - \bar{v}\|_{\infty} \cdot 1$$

$$\Rightarrow \|\hat{T}^{\pi} v - \hat{T}^{\pi} \bar{v}\|_{\infty} \leq \beta \|v - \bar{v}\|_{\infty}$$

where, $\beta = \gamma^{\pi}(s)$

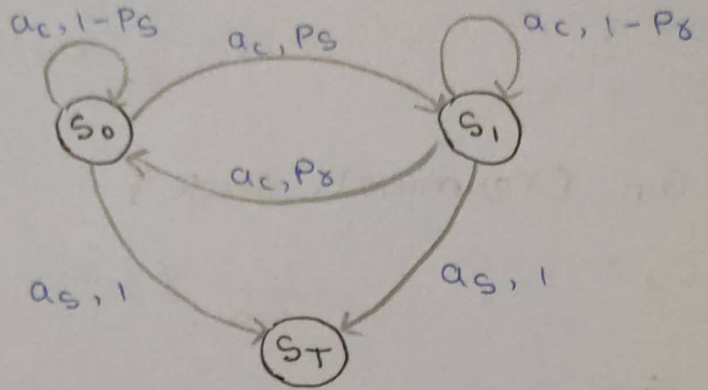
$\therefore$ The given operator is a contraction given $0 < \gamma^{\pi}(s) < 1$

Contraction factor $= \gamma^{\pi}(s)$

Hence Proved

Question 2 a

Transition Probabilities for the given MDP



$P(S_0 | S_0, a_s) = 0$

$P(S_1 | S_0, a_s) = 0$

$P(S_T | S_0, a_s) = 1$

$P(S_0 | S_1, a_s) = 0$

$P(S_1 | S_1, a_s) = 0$

$P(S_T | S_1, a_s) = 1$

$P(S_0 | S_0, a_c) = 1 - P_s$

$P(S_1 | S_0, a_c) = P_s$

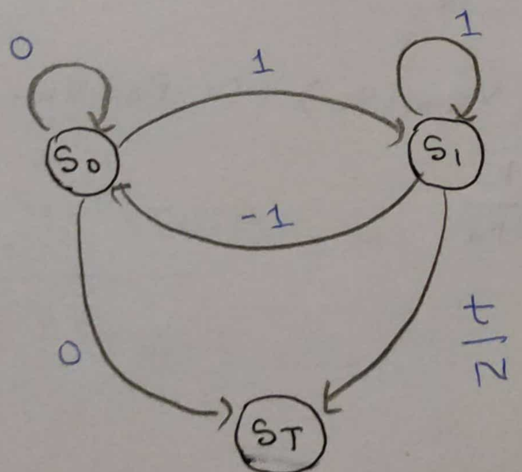
$P(S_T | S_0, a_c) = 0$

$P(S_0 | S_1, a_c) = P_r$

$P(S_1 | S_1, a_c) = 1 - P_r$

$P(S_T | S_1, a_c) = 0$

Question 2 b



$t = 1, 2, \ldots N$

Question 2 c

By Dynamic Programming,

$$V_K(S) = \max_a \sum_{s', r} P(s', r \mid s, a)\left[r + V_{K+1}(s')\right]$$

For state S_T [Terminal state]

$$V_K(S_T) = 0 \qquad , \forall K = 1, \ldots N$$

For state S_0

$$V_K(S_0) = \max \begin{cases} P_S\left[1 + V_{K+1}(S_1)\right] + (1-P_S)\left[0 + V_{K+1}(S_0)\right], \\ 1\left[0 + V_{K+1}(S_T)\right] \end{cases}$$

$$= \max\left\{P_S + P_S V_{K+1}(S_1) + (1-P_S) V_{K+1}(S_0) \;,\; 0\right\}$$

For state S_1

$$V_K(S_1) = \max \begin{cases} P_r\left[-1 + V_{K+1}(S_0)\right] + (1-P_r)\left\{1 + V_{K+1}(S_1)\right], \\ 1\left[\frac{K}{N} + V_{K+1}(S_T)\right) \end{cases}$$

$$= \max \begin{cases} P_r V_{K+1}(S_0) + (1-P_r) V_{K+1}(S_1) + (1-2P_r), \\ \frac{K}{N} \end{cases}$$

Question ③ ⓐ

To show: Maximizing the expected total sum of rewards $\mathbb{E}[\sum_{t=1}^{T} x_{t,I_t}]$ is equivalent to minimizing the regret

$$\Rightarrow \max \mathbb{E}\left[\sum_{t=1}^{T} x_{t,I_t}\right] = \min\left(-\mathbb{E}\left[\sum_{t=1}^{T} x_{t,I_t}\right]\right)$$

$$= \min\left[T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} x_{t,I_t}\right]\right]$$

$$= \min\left[T\mu^* - \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{T} x_{ti}\, \mathbb{1}\{I_s = i\}\right]\right]$$

Dividing and Multiplying by $\sum_{t=1}^{T} \mathbb{1}\{I_s = i\}$

$$= \min\left(T\mu^* - \mathbb{E}\left[\sum_{i=1}^{K}\left[\frac{\sum_{t=1}^{T} x_{ti}\, \mathbb{1}\{I_s=i\}}{\sum_{t=1}^{T} \mathbb{1}\{I_s=i\}} \cdot \sum_{t=1}^{T} \mathbb{1}\{I_s=i\}\right]\right]\right)$$

Here, $\frac{\sum_{t=1}^{T} x_{ti}\, \mathbb{1}\{I_s=i\}}{\sum_{\tau=1}^{T} \mathbb{1}\{I_s=i\}} = \mu_i$ and $\sum_{t=1}^{T} \mathbb{1}\{I_s=i\} = N_i(T)$

$$\therefore = \min\left(T\mu^* - \mathbb{E}\left[\sum_{i=1}^{K} \mu_i . N_i(T)\right]\right)$$

$$= \min(\text{Regret}(T))$$

Hence shown that $\max \mathbb{E}\left[\sum_{t=1}^{T} x_{t,I_t}\right] = \min \text{Regret}(T)$

Question (3) (b) (i)

Given: $K = 2$

Rewards are bounded in $[0, 1]$

Arm 1 is optimal $\Rightarrow \mu^* = \mu_1$

$\Delta = \mu_1 - \mu_2$

To prove: $\text{Regret}(T) \leq m\Delta + \Delta(T-2m)\mathbb{E}[\mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}]$

Proof:

$$\text{Regret}(T) = T\mu^* - \mathbb{E}\left[\sum_{i=1}^{K} \mu_i N_i(T)\right]$$

Phase 1:

In phase 1 for time $t = 1, 2, \ldots 2m$ the 2 arms are played in a round-robin fashion; i.e each arm is played m times $\left(\because \frac{t}{2} = \frac{2m}{2} = m\right)$

$$\therefore \text{Regret}(T_1) = (2m)\mu_1 - \mathbb{E}\left[\sum_{i=1}^{2} \mu_i N_i(T)\right]$$

$$= (2m)\mu_1 - \mathbb{E}\left[\mu_1 N_1(T) + \mu_2 N_2(T)\right]$$

Here, $N_1(T) = N_2(T) = m$

$$= 2m\mu_1 - \mathbb{E}[\mu_1 \cdot m + \mu_2 m]$$

$$= 2m\mu_1 - \mu_1 m - \mu_2 m$$

$$= m\mu_1 - m\mu_2$$

$$= m(\mu_1 - \mu_2)$$

$$\text{regret}(T_1) = m\Delta \longrightarrow (1)$$

$\therefore$ Regret in phase 1 is $m\Delta$

Phase 2:

In phase 2 for time $t = 2m+1, 2m+2, \ldots T$, the arm with the best sample mean till $t = 2m$ is played.

$\therefore$ Total time $T_2 = T - (2m+1) - 1$

$= T - 2m$

Number of times arm 1 is played:

$$N_1(T_2) = T_2 * \mathbb{1}_{\hat{\mu}_1(2m) > \hat{\mu}_2(2m)}$$

$$= (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) > \hat{\mu}_2(2m)}$$

Number of times arm 2 is played:

$$N_2(T_2) = T_2 * \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}$$

$$= (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}$$

$$\therefore \text{Regret}(T_2) = (T-2m)\mu_1 - \mathbb{E}\Big[\mu_1 (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) > \hat{\mu}_2(2m)} + \mu_2 (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\Big]$$

$$= (T-2m)\mu_1 - \mathbb{E}\Big[\mu_1 (T-2m)\big(1 - \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\big) + \mu_2 (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\Big]$$

$$= (T-2m)\mu_1 - \mathbb{E}\Big[\mu_1 (T-2m) - \mu_1 (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)} + \mu_2 (T-2m)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\Big]$$

$$= (T-2m)\mu_1 - \mu_1 (T-2m) - \mathbb{E}\Big[(\mu_2 - \mu_1)\, \mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)} \cdot (T-2m)\Big]$$

Cancelling out $(T-2m)\mu_1$

$$\leq -\mathbb{E}\left[-\Delta(T-2m)\,\mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\right]$$

$$\leq \Delta(T-2m)\,\mathbb{E}\left[\mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\right]$$

$\therefore$ Regret in Phase 2 is

$$\text{Regret}(T_2) \leq \Delta(T-2m)\,\mathbb{E}\left[\mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\right] \rightarrow (2)$$

From (1) & (2)

Total Regret

$$\text{Reget}(T) \leq \text{Regret}(T_1) + \text{Regret}(T_2)$$

$$\leq m\Delta + \Delta(T-2m)\,\mathbb{E}\left[\mathbb{1}_{\hat{\mu}_1(2m) \leq \hat{\mu}_2(2m)}\right]$$

Hence Proved

Question 3 b ii

$$\text{Regret}(T) \le m\Delta + \Delta(T-2m)\,\mathbb{E}[\mathbb{1}_{\hat{\mu}_1(2m) \le \hat{\mu}_2(2m)}]$$

we know that,

$$\mathbb{E}[\mathbb{1}_A] = P(A)$$

$$\therefore \text{Regret}(T) \le m\Delta + \Delta(T-2m)\,P(\hat{\mu}_1(2m) \le \hat{\mu}_2(2m)) \rightarrow 3$$

Here, considering $P(\hat{\mu}_1(2m) \le \hat{\mu}_2(2m))$

Expanding $\hat{\mu}_1(2m)$ and $\hat{\mu}_2(2m)$

$$P\left(\frac{\sum_{s=1}^{2m} X_{s_1} \mathbb{1}\{I_s = 1\}}{\sum_{s=1}^{2m} \mathbb{1}\{I_s = 1\}} \le \frac{\sum_{s=1}^{2m} X_{s_2} \mathbb{1}\{I_s = 2\}}{\sum_{s=1}^{2m} \mathbb{1}\{I_s = 2\}} \right)$$

Here, $\sum_{s=1}^{2m} \mathbb{1}\{I_s = 1\} = N_1(2m) = m$

$\sum_{s=1}^{2m} \mathbb{1}\{I_s = 2\} = N_2(2m) = m$ (Phase 1)

$$\Rightarrow P\left(\frac{\sum_{s=1}^{2m} X_{s_1} \mathbb{1}\{I_s = 1\}}{m} \le \frac{\sum_{s=1}^{2m} X_{s_2} \mathbb{1}\{I_s = 2\}}{m} \right)$$

Adding and substracting $\frac{\mu_1}{m}$ on LHS and $\frac{\mu_2}{m}$ on RHS

$$\Rightarrow P\left(\frac{\sum_{s=1}^{2m} X_{s_1} \mathbb{1}\{I_s = 1\} - \mu_1}{m} + \frac{\mu_1}{m} \le \frac{\sum_{s=1}^{2m} X_{s_2} \mathbb{1}\{I_s = 2\} - \mu_2}{m} + \frac{\mu_2}{m} \right)$$

Rearranging the terms

$$P\left(\frac{\mu_1-\mu_2}{m} \le \frac{\left[\sum_{s=1}^{2m} x_{s_2}\mathbb{1}_{\{I_s=2\}}-\mu_2\right]}{m} - \frac{\left[\sum_{s=1}^{2m} x_{s_1}\mathbb{1}_{\{I_s=1\}}-\mu_1\right]}{m}\right)$$

w.k.t $E[x_1]=\mu_1$ and $E[x_2]=\mu_2$

$$P\left(\frac{\Delta}{m} \le \frac{1}{m}\left(\left(\sum_{s=1}^{2m} x_{s_2}\mathbb{1}_{\{I_s=2\}}-\mu_2\right)+\left(\sum_{s=1}^{2m} x_{s_1}\mathbb{1}_{\{I_s=1\}}-\mu_1\right)\right)\right)$$

$$\Rightarrow P\left(\frac{\Delta}{m} \le \frac{1}{m}\left(\left(\sum_{s=1}^{2m} x_{s_2}\mathbb{1}_{\{I_s=2\}}-E[x_2]\right)+\left(\sum_{s=1}^{2m} x_{s_1}\mathbb{1}_{\{I_s=1\}}-E[x_1]\right)\right)\right)$$

$$\Rightarrow P\left(\frac{\Delta}{m} \le \frac{1}{m}\left(\sum_{s=1}^{2m} x_{s,I_s}-E[x_{I_s}]\right)\right)$$

Dividing by 2 on both sides

$$\Rightarrow P\left(\frac{\Delta}{2m} \le \frac{1}{2m}\left(\sum_{s=1}^{2m} x_{s,I_s}-E[x_{I_s}]\right)\right)$$

Here, as $\{x_{s,I_s}\}_s$ is i.i.d and $x_{s,I_s} \in [0,1]$

$\therefore \Rightarrow a=0$, $b=1$ and

$\epsilon = \frac{\Delta}{2m}$, $n=2m$

$\therefore$ using Chernoff-Hoefding's bound,

$$\Rightarrow P\left(\frac{\Delta}{2m} \le \frac{1}{2m}\left(\sum_{s=1}^{2m} x_{s,I_s}-E[x_{I_s}]\right)\right) \le e^{-2*2m*\frac{\Delta^2}{4m^2}/(1-0)^2}$$

$$\le e^{-\left(\frac{1}{m^2}\right)m\Delta^2}$$

$$\leq e^{-cm\Delta^2} \quad , \text{ where } c > 0$$

Substituting in (3)

$$\text{Regret}(T) \leq m\Delta + \Delta(T-2m)e^{-cm\Delta^2}$$

$$\leq m\Delta + \Delta T e^{-cm\Delta^2}$$

Hence shown, that $\text{Regret}(T) \leq m\Delta + \Delta T e^{-cm\Delta^2}$.

Let $m = \dfrac{\log T}{c\Delta^2}$

then, $\text{Regret}(T) \leq \dfrac{\log T}{c\Delta^2} . \Delta + \Delta(T)\cancel{-2}\, e^{-c \cdot \frac{\log T}{c\Delta^2} \cdot \Delta}$

$$\leq \frac{\log T}{c\Delta} + \Delta T e^{-\log T}$$

$$\leq \frac{\log T}{c\Delta} + \Delta T e^{\log T^{-1}}$$

$$\leq \frac{\log T}{c\Delta} + \Delta T . T^{-1}$$

$$\leq \frac{\log T}{c\Delta} + \Delta$$

$$\text{Regret}(T) = O\left(\frac{1}{\Delta}\log T\right)$$

Question 3 b iii

Difficulty in implementing the algorithm to obtain a sub-linear regret:

i) It takes linear time on T to compute μ^*.

ii) It is difficult to compute the expected reward obtained by the algorithm.