Q1. (a)

$$T^* Q(s,a) = r(s,a) + \gamma \int P(s'|s,a) \max_{a' \in A} Q(s',a') \, ds'$$

Consider Q_1 and Q_2

$$|T^*Q_1(s,a) - T^*Q_2(s,a)| = \gamma \left| \int \left(\max_{a' \in A} Q_1(s',a') - \max_{a' \in A} Q_2(s',a') \right) P(s'|s,a) \, ds' \right|$$

$$\leq \gamma \int \left| \max_{a' \in A} Q_1(s',a') - \max_{a' \in A} Q_2(s',a') \right| P(s'|s,a) \, ds'$$

$$\leq \gamma \int \sup_{s' \in S} \max_{a' \in A} |Q_1(s',a') - Q_2(s',a')| \, P(s'|s,a) \, ds'$$

$$\leq \gamma \sup_{(s',a') \in S \times A} |Q_1(s',a') - Q_2(s',a')| \int P(s'|s,a) \, ds'$$

$$= \gamma \| Q_1 - Q_2 \|_\infty$$

$$\begin{aligned}
\| Q_k - Q^* \|_\infty &= \| T Q_{k-1} - Q^* \|_\infty \\
&\leq \| T Q_{k-1} - T Q^* \|_\infty \\
&\leq \gamma \| Q_{k-1} - Q^* \|_\infty \\
&\vdots \\
&\leq \gamma^k \| Q_0 - Q^* \|_\infty \\
&= \left(1 - (1 - \gamma) \right)^k \| Q_0 - Q^* \|_\infty
\end{aligned}$$

$$\leq e^{-(1-\gamma)k} \|Q_0 - Q^*\|_\infty \qquad \left(1-x \leq e^{-x} \quad \forall x \in (0,1) \right)$$

We want

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

which can be ensured by setting

$$e^{-(1-\gamma)k} \|Q_0 - Q^*\|_\infty \leq \epsilon$$

$$\text{or} \qquad k \geq \frac{1}{1-\gamma} \log\left(\frac{\|Q_0 - Q^*\|_\infty}{\epsilon}\right)$$

Let $Q_0 \equiv 0$.

Using the fact that $\sum_{t=1}^{\infty} \gamma^{t-1} r(s,a) \leq \frac{1}{1-\gamma}$

$$\|Q^*\|_\infty \leq \frac{1}{1-\gamma}$$

$$\text{Sample Complexity} = O\left(\frac{1}{1-\gamma}\log\left(\frac{1-\gamma}{\epsilon}\right)\right)$$

(b)

Policy iteration

Policy Evaluation : Compute Q^{π_k}

Policy Improvement : $\pi_{k+1}(s) \leftarrow \arg\max_a Q^{\pi_k}(s,a)$

$$Q^{\pi_k} \leq T^* Q^{\pi_k} \leq Q^{\pi_{k+1}}$$

$$Q^{\pi_k} \leq Q^{\pi_{k+1}}$$

$$V^{\pi_k} \leq V^{\pi_{k+1}}$$

$$\int \pi_k(a|s)\, Q^{\pi_k}(s,a) \leq \int \pi_{k+1}(a|s)\, Q^{\pi_{k+1}}(s,a)$$

$$Q^{\pi_{k+1}}(s,a) = r(s,a) + \gamma \int P(s'|s,a)\, \pi_{k+1}(a'|s')\, Q^{\pi_{k+1}}(s,a)\, ds'da'$$

$$\geq r(s,a) + \gamma \int P(s'|s,a)\, Q^{\pi_{k+1}}\left(s, \underset{a}{\operatorname{argmax}}\, Q^{\pi_k}(s,a)\right) ds'da'$$

$$\geq r(s,a) + \gamma \int P(s'|s,a) \max_a Q^{\pi_k}(s,a)$$

$$= T^* Q^{\pi_k}(s,a)$$

$$Q^{\pi_{k+1}} \geq T^* Q^{\pi_k}$$

$$Q^{\pi_k} \geq T^* Q^{\pi_{k-1}}$$

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

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

$$\vdots$$

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

Use same analysis as in part (a).

(c) $\tilde{\pi} = (\pi_1, \pi_2, \ldots)$

$$V^{\tilde{\pi}} = \liminf_{t \to \infty} T^{\pi_1} T^{\pi_2} \ldots\ldots T^{\pi_t} V_0$$

$$V^+ = \sup_{\tilde{\pi}'} V^{\tilde{\pi}}$$

$$V^* = TV^*$$

$$V^* = V^{\pi^*}$$
$$= T^{\pi^*} T^{\pi^*} \cdots\cdots T^{\pi^*} \cdots V^0$$

$$\therefore \quad \pi^* \in \{ \tilde{\pi} : \tilde{\pi} = (\pi_1, \pi_2, \ldots\ldots) \}$$

$$\Rightarrow \quad V^{\pi^*} \leq \sup_{\tilde{\pi}} V^{\tilde{\pi}} = V^+ \qquad \text{——— } (1)$$

Consider an arbitrary non-stationary policy
$\tilde{\pi} = (\pi_1, \pi_2, \ldots\ldots)$

$$T^{\pi_k} V_0 \leq T^* V_0$$

because T^* is optimal

$$T^{\pi_1} T^{\pi_2} \dots T^{\pi_k} V_0 \le (T^*)^k V_0$$

$$\liminf_{k\to\infty} T^{\pi_1} T^{\pi_2} \dots T^{\pi_k} V_0 \le \lim_{k\to\infty} (T^*)^k V_0$$

$$= V^*$$

Since $\tilde{\pi}$ was arbitrary,

$$V^{\dagger} = \sup_{\tilde{\pi}} V^{\tilde{\pi}} \le V^* \quad \text{——— } (2)$$

from (1) and (2),

$$V^* = V^{\dagger}.$$

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

<u>Monotonic</u>

Let $V_1(s) \leq V_2(s) \quad \forall s \in \mathcal{S}$

$$(\hat{T}^{\pi} V_1)(s) = \gamma^{\pi}(s) \int P(s'|s) V_1(s') \, ds'$$

$$\overset{?}{\leq} \gamma^{\pi}(s) \int P(s'|s) V_2(s') \, ds'$$

This is not true when $\gamma^{\pi}(s) < 0$.

Under the assumption that

$$r(s,a) \geq 0 \quad \forall s, \in \mathcal{S},$$

$\hat{T}$ is monotonic.

<u>Contraction</u> :

$$\left| \hat{T}^{\pi} V_1(s) - \hat{T}^{\pi} V_2(s) \right| \leq r^{\pi}(s) \int P(s'|s) \left| V_1(s') - V_2(s') \right| ds' \quad \forall s$$

$$\leq r^{\pi}(s) \sup_{s \in \mathcal{S}} \left| V_1(s) - V_2(s) \right| \int P(s'|s)\, ds'$$

$$\| \hat{T}^{\pi} V_1 - \hat{T}^{\pi} V_2 \|_{\infty} \leq r^{\pi}(s) \, \| V_1 - V_2 \|_{\infty}$$

$\hat{T}$ is not a contraction when $\gamma^{\pi}(s)$.

When, $\gamma(s,a) \in (0,1)$,

then $\hat{T}$ is a contraction.

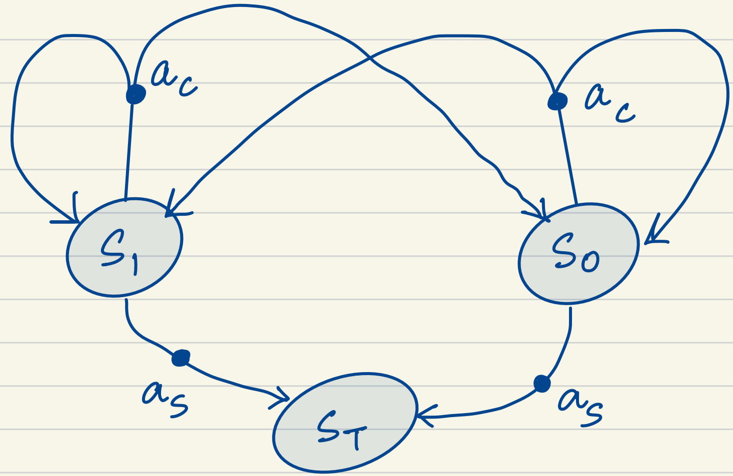
Q2.

Relative rank : The rank of the current item when compared with all the previously seen items.

Should I stop when relative rank > 1 — No.

S_1 : the relative rank of the current item is 1.

S_0 : " " " " " " " is not 1.



$$P\left(s_{k+1} = s_1 \mid s_k = s_1 , a_c\right) = \frac{1}{k+1}$$

$$P\left(s_{k+1} = s_0 \mid s_k = s_1 , a_c\right) = \frac{k}{k+1}$$

Similarly for $s_k = s_0$.

$$P(s_{k+1} = s_T \mid s_k = s_1, a_s) = 1$$

$$\quad " \qquad\qquad a_c = 0$$

Similarly for $s_k = s_0$

(b)

$$\text{Return} = \mathbb{E}\left(\sum_{k=0}^{N-1} c_k(s_k, a_k) + c_k(x_N).\right)$$

you want this to correspond to selecting the best candidate.

$$c_k(s_k, a_k) = \begin{cases} \frac{k}{N} & \text{if } x_k = s_1 \text{ and } a_k = a_s \\ 0 & \text{otherwise} \end{cases}$$

$$C_N(S_N) = \begin{cases} 1 & S_N = 1 \\ 0 & S_N = 0 \text{ or } S_N = T \end{cases}$$

Why $\frac{k}{N}$?

$$P(k^{th} \text{ candidate} = \text{Top ranked candidate})$$

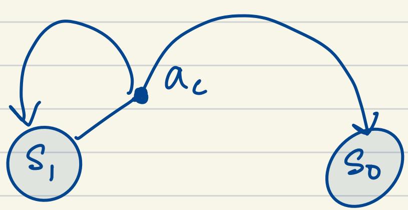
$$= \frac{{}^{N-1}C_{k-1}}{{}^{N}C_{k}} = \frac{k}{N}$$

(C) $$V_N^*(s_N) = c_N(s_N)$$

$$V_k^*(s_k) = \max_{a_k} \mathbb{E}\left(c_k(s_k, a_k) + V_{k+1}^*(s_{k+1}) \right)$$

$$V_k^*(s_1) = \max_{a_k \in \{a_s, a_c\}} \mathbb{E}\left(c_k(s_k = s_1, a_k) + V_{k+1}^*(s_{k+1}) \right)$$

$$= \max\left(\frac{k}{N} + V_{k+1}^*(s_T) \;,\; \frac{k}{k+1} V_{k+1}^*(s_0) + \frac{1}{k+1} V_{k+1}^*(s_1) \right)$$



$$V^*_k(s_0) = \max_{a_k} \mathbb{E}\left(c_k(s_k = s_0, a_k) + V^*_{k+1}(s_{k+1}) \right)$$

$$= \max\left(0 + V^*_{k+1}(s_T) \,,\ \frac{k}{k+1} V^*_{k+1}(s_0) + \frac{1}{k+1} V^*_{k+1}(s_1) \right)$$

$$= \frac{k}{k+1} V^*_{k+1}(s_0) + \frac{1}{k+1} V^*_{k+1}(s_1)$$

Q3.

(a)

$$\mathbb{E}\left(\sum_{t=1}^{T} X_{t,I_t}\right) = \mathbb{E}\left(\sum_{i=1}^{k} \mu_i \, N_i(T)\right)$$

$$X_{1,I_1}, \; X_{2,I_2}, \; \ldots$$

$$X_{1,2}, \; X_{2,1}, \; X_{3,1}, \; X_{4,2}, \; \ldots$$

$$N_i(T) = \sum_{i=1}^{T} \mathbb{1}_{\{I_t = i\}}$$

$$\equiv \max \; \mathbb{E} \sum_{i=1}^{k} \mu_i N_i(T) \; - \; T\mu^*$$

$$\equiv \min \; T\mu^* \;\; \mathbb{E} \sum_{i=1}^{k} \mu_i N_i(T)$$

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

$$\equiv \min \; \sum_{i=1}^{k} \underbrace{(\mu^* - \mu_i)}_{\Delta_i} \, \mathbb{E}\, N_i(T)$$

$$\min \underbrace{\sum_{i=1}^{k} \Delta_i \, \mathbb{E} N_i(T)}_{\text{Regret}(T)}$$

(b) (i)

$$\text{Regret}(T) = \Delta_1 \mathbb{E} N_1(T) + \Delta_2 \mathbb{E} N_2(T)$$

$$= \Delta \mathbb{E} N_2(T)$$

$$= \Delta \mathbb{E}\left(N_2(mk)\right) + \Delta \mathbb{E}\left(N_2(T) - N_2(mk)\right)$$

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

$$\mathbb{E}\left[\mathbb{1}_{\{\hat{\mu}_1(2m) \le \hat{\mu}_2(2m)\}}\right] = \mathbb{P}\{\hat{\mu}_1(2m) \le \hat{\mu}_2(2m)\}$$

$$= P\left\{ \frac{1}{m}\sum_{i=1}^{m} X_{2i-1,1} \le \frac{1}{m}\sum_{i=1}^{m} X_{2i,2} \right\}$$

$$= P\left\{ \frac{1}{m}\sum_{i=1}^{m} \left(X_{2i-1,1} - X_{2i,2}\right) \le 0 \right\}$$

$$= P\left\{ \frac{1}{m}\sum_{i=1}^{m} \left(\underbrace{X_{2i-1,1} - X_{2i,2}}_{\in[-1,1]} - \Delta\right) \le -\Delta \right\}$$

$$\leq e^{-2m\Delta^2/4}$$

$$\leq e^{-m\Delta^2/2}$$

$$\text{Regret}(T) \leq m\Delta + \Delta T e^{-m\Delta^2/8}$$

Choose m such that $\Delta T e^{-m\Delta^2/2} = 1$

$$m^* = \frac{2}{\Delta^2} \log(\Delta T)$$

$$\text{Regret}(T) \leq \frac{2}{\Delta} \log(\Delta T)$$

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

(iii) m^* depends on Δ, which is typically unknown.