Homework 4

Problem 1 For the multi-armed bandits problem, denote X_t as the (stochastic) reward we obtain at time t. For action $a \in \mathbb{A}$, where $\mathbb{A}$ is the set of all actions, denote μ_a as the mean reward associated with action a. Also denote $T_a(n)$ as the number of times action a has been chosen, for a total game duration of n. Assume that μ_* is the optimal mean reward. We know that the regret can be computed as:

$$R_n = \sum_{t=1}^{n} \left(\mu_* - \mathbb{E}[X_t]\right).$$

Prove that it can also be computed as:

$$R_n = \sum_{a \in \mathbb{A}} (\mu_* - \mu_a) \cdot \mathbb{E}[T_a(n)].$$

Problem 2 For the Markov decision process problem over a time horizon of n, we have discussed in the class the value iteration method to find the optimal policy $\pi^* = \{\pi_t^*(a_t|s_t)\}_{t=1}^n$ given knowledge about the Markov transition probabilities at each time point $\{P_t(s_{t+1}|s_t, a_t)\}_{t+1}^n$. It iterates based on the Bellman optimality equation as follows:

$$\begin{aligned}
&\text{Initialize } V_{n+1}(s) = 0, \forall s \in \mathbb{S} \\
&\text{For } t = n, n-1, \ldots, 1 \\
&\qquad Q_t(s,a) = r_t(s,a) + \sum_{s \in \mathbb{S}} P_t(s_{t+1} = s' | s_t = s, a_t = a) \cdot V_{t+1}(s') \\
&\qquad V_t(s) = \max_{a \in \mathbb{A}} Q_t(s,a)
\end{aligned}$$

We can also use the following policy iteration method by maintaining and updating the policy $\pi = \{\pi_t\}_{t=1}^n$.

$$\begin{aligned}
&\text{Initialize } \pi^{(0)} = \left\{\pi_t^{(0)}\right\}_{t=1}^n \text{ randomly} \\
&\text{For } k = 1, 2, \ldots \\
&\qquad \text{Evaluate } Q_t^{\pi^{(k-1)}}(s,a), \forall s \in \mathbb{S}, a \in \mathbb{A}, t = 1, \ldots, n, \text{ using the Bellman equation (1),(2)} \\
&\qquad \pi_t^{(k)}(s) = \arg\max_{a \in \mathbb{A}} Q_t^{\pi^{(k-1)}}(s,a), \forall s \in \mathbb{S}, t = 1, \ldots, n
\end{aligned}$$

Question: after how many iterations K (if any), would $\pi^{(K)}$ in the policy iteration algorithm become the optimal policy? Prove your result.

The Bellman equation:

$$Q_t(s, a) = r_t(s, a) + \sum_{s \in \mathbb{S}} P_t(s_{t+1} = s' | s_t = s, a_t = a) \cdot V_{t+1}(s') \tag{1}$$

$$V_t(s) = \sum_{a \in \mathbb{A}} \pi_t(a|s) \cdot Q_t(s, a) \tag{2}$$

Hint: prove inductively that $\forall t \geq n - k$, $Q_t^{\pi^{(k)}} = Q_t^{\pi^*}$.

Machine Learning
Homework - 4

Problem 1

$\boxed{\text{Some terms}}$: X_t : Reward at time t

Action $a \in A$, μ_a : mean reward of action a

$T_a(n)$: number of times a has been chosen

μ_t : optimal Reward

Regret $$R_n = \sum_{t=1}^{n} (\mu_* - E[x_t]) \quad - (1)$$

To Prove : $$R_n = \sum_{a \in A} (\mu_* - \mu_a) \cdot E[T_a(n)] \quad - (2)$$

from (1)

$$R_n = \sum_{t=1}^{n} (\mu_* - E[x_t])$$; x_t is i.i.d (doesn't depend on t)

$$= n\mu_* - E[\sum_{t=1}^{n} x_t]$$

Now, $$\sum_{t=1}^{n} x_t = \sum_{t}^{n} \sum_{a}^{A} x_t \cdot \mathbb{1}\{A_t = a\}$$ ⇒ Sum the reward obtained when action a happend

$$\rightarrow R_n = n\mu_* - E[\sum_{t=1}^{n} x_t] = \sum_{a}^{A} \sum_{t}^{n} E[(\mu_* - x_t) \cdot \mathbb{1}\{A_t * a\}]$$

Expected reward in round t conditioned on A_t is

$\mu_{A_t} = \mu_a$ ← action at that t

$$\rightarrow \mathbb{E}\left[(\mu_* - X_t)\,\mathbb{1}\{A_t = a\} \mid A_t\right] = \mathbb{1}\{A_t = a\} \cdot \mathbb{E}\left[(\mu_* - X_t) \mid A_t\right]$$

$$= \mathbb{1}\{A_t = a\}\,(\mu_* - \mu_{a_t})$$

$$= \mathbb{1}\{A_t = a\}\,(\mu_* - \mu_a)$$

$$\rightarrow R_n = \sum_a^A \sum_t^n \left(\mathbb{1}\{A_t = a\}(\mu_* - \mu_a)\right) = \sum_a^A \left((\mu_* - \mu_a) \underbrace{\sum_t^n \mathbb{1}\{A_t = a\}}_{\mathbb{E}[T_a(n)]}\right)$$

$$\rightarrow \underline{R_n = \sum_a^A (\mu_* - \mu_a) \cdot \mathbb{E}[T_a(n)]}$$

Hence Proved

Problem 2

Proof:

Initialize $\pi^{(0)} = \{\pi_t^{(0)}\}_{t=1}^{n}$ randomly

for $k = 1, 2, \ldots$

Now, for $t = n, n-1, \ldots, 1$ $\Rightarrow$

$$Q_t^{\pi(k-1)}(s,a) = r_t(s,a) + \sum_{s \in S} P_t(s_{t+1} = s' \mid s_t = s, a_t = a) \cdot V_{t+1}^{\pi(k-1)}(s')$$

$$V_t^{\pi(k-1)}(s) = \sum_{a \in A} \pi_t^{(k-1)}(a|s) \cdot Q_t^{\pi(k-1)}(s,a)$$

$$\pi_t^{(k)}(s) = \arg\max_{a \in A} Q_t^{\pi(k-1)}(s,a), \ \forall s \in S, t = 1, \ldots, n$$

We want to prove $\forall t \geq n-k$, $Q_t^{\pi(k)} = Q_t^{\pi^*}$ by induction

1. When $t = n$

$$Q_n^{\pi(k)}(s,a) = r_a(s,a) = Q_n^{\pi^*}(s,a)$$

holds for all $k \geq 0$

2. If $Q_{t+1}^{\pi(k)} = Q_{t+1}^{\pi^*}$, $\forall k \geq n-(t+1)$ holds for $t+1$,

then for t: $\forall k \geq n-t$,

$$Q_t^{\pi^{(k)}}(s,a) = r_t(s,a) + \sum_{s\in S} P_t\left(S_{t+1}=s' \,\middle|\, \begin{matrix} S_t = s \\ a_t = a\end{matrix}\right) \cdot V_{t+1}^{\pi^{(k)}}(s')$$

$$= r_t(s,a) + \sum_{s\in S} P_t\left(S_{t+1}=s' \,\middle|\, \begin{matrix} S_t = s \\ a_t = a\end{matrix}\right)\left(\sum_{a\in A} \pi_{t+1}^{(k)}(a|s') \cdot Q_{t+1}^{\pi^{(k)}}(s',a)\right)$$

$$= r_t(s,a) + \sum_{s\in S} P_t\left(S_{t+1}=s' \,\middle|\, \begin{matrix} S_t = s \\ a_t = a\end{matrix}\right)$$

$$\left(\sum_{a\in A} \mathbb{1}\{a = \arg\max_a Q_{t+1}^{\pi^{(k)}}(s',a)\}\, Q_{t+1}^{\pi^{(k)}}(s',a)\right)$$

$$= r_t(s,a) + \sum_{s\in S} P_t\left(S_{t+1}=s' \,\middle|\, \begin{matrix} S_t = s \\ a_t = a\end{matrix}\right) \max_a Q_{t+1}^{\pi^{(k)}}(s',a)$$

$$= r_t(s,a) + \sum_{s\in S} P_t\left(S_{t+1}=s' \,\middle|\, \begin{matrix} S_t = s \\ a_t = a\end{matrix}\right) \max_a Q_{t+1}^{\pi^*}(s',a)$$

$$= Q_t^{\pi^*}(s,a)$$

So we proved that $Q_t^{\pi^{(k)}} = Q_t^{\pi^*}, \forall\, t \geq n-k$

So

$$Q_t^{\pi(n-1)} = Q_t^{\pi^*} \quad \text{for } t = 1, 2, \cdots, n$$

$$\Rightarrow \pi_t^{(n)}(s) = \underset{a \in A}{\operatorname{argmax}}\, Q_t^{\pi(n-1)}(s, a)$$

$$= \underset{a \in A}{\operatorname{argmax}}\, Q_t^{\pi^*}(s, a)$$

$$= \pi_t^*(s)$$

$$\Rightarrow \pi^{(n)} = \pi^* \qquad \left[k = \log \left(\underset{a \in A}{\arg\max}\, Q_t^{\pi^*}(s, a) \right] \right.$$

$$\therefore \boxed{k = n}$$