COMP 767 (Reinforcement Learning)
Assignment 1

Arna Ghosh (260748358), Arnab Kumar Mondal (260906419)

29th January 2020

1 Theory Part

Answer 1. We have $\mu_i = \mathbb{E}[R_i]$ and $\mu^* = \max_{i=1}^{K} \mu_i$ where K is the number of arms. Let us denote $\bar{\mu}_i$ be the estimate of i-th arm such that $\bar{\mu}_i = \sum_{t=1}^{T/K} \frac{R_{it}}{T/K} = \sum_{t=1}^{T/K} \frac{K}{T} R_{it}$ where R_{it} is the random variable denoting the reward sample obtained from t-th trial of the i-th arm. If we take $n = T/K$ then we can define random variable $\bar{\mu}_i = \frac{\sum_{t=1}^{n} R_{it}}{n}$. By Hoeffding's inequality we have:

$$\mathbb{P}[|\bar{\mu}_i - \mathbb{E}[\bar{\mu}_i]| \geq \lambda] \leq 2e^{-2n\lambda^2}$$

where $\lambda \geq 0$. We also have:

$$\mathbb{E}[\bar{\mu}_i] = \mathbb{E}[\frac{\sum_{t=1}^{n} R_{it}}{n}] = \frac{\sum_{t=1}^{n} \mathbb{E}[R_{it}]}{n} = \frac{\sum_{t=1}^{n} \mu_i}{n} = \mu_i$$

$$\implies \mathbb{P}[|\bar{\mu}_i - \mu_i]| \geq \lambda] \leq 2e^{-2n\lambda^2}$$

Now let E_i be the event of $|\bar{\mu}_i - \mu_i| < \lambda$, then $\mathbb{P}[\Omega \setminus E_k] = \mathbb{P}[|\bar{\mu}_i - \mu_i]| \geq \lambda] \leq 2e^{-2n\lambda^2}$ where Ω is the universal set.

$$\mathbb{P}[\bigcap_{i=1}^{K} E_i] = \mathbb{P}[\Omega \setminus (\bigcup_{i=1}^{K} \Omega \setminus E_i)] = 1 - \mathbb{P}[\bigcup_{i=1}^{K} \Omega \setminus E_i]$$

By Union bound we have:

$$\mathbb{P}[\bigcup_{i=1}^{K} \Omega \setminus E_i] \leq \sum_{i=1}^{K} \mathbb{P}[\Omega \setminus E_i]$$

$$\implies \mathbb{P}[\bigcap_{i=1}^{K} E_i] \geq 1 - \sum_{i=1}^{K} \mathbb{P}[\Omega \setminus E_i] \geq 1 - \sum_{i=1}^{K} 2e^{-2n\lambda^2}$$

$$\implies \mathbb{P}[\bigcap_{i=1}^{K} E_i] \geq 1 - 2Ke^{-2n\lambda^2}$$

$\bigcap_{i=1}^{K} E_i$ implies that $|\bar{\mu}_i - \mu_i| < \lambda \quad \forall i \in \{1, 2, .., K\}$. Let $\hat{i} = \arg\max \bar{\mu}_i$ and if $\hat{i}$ is the optimal arm chosen then $|\mu^* - \mu_{\hat{i}}| = 0$ else we have two inequalities $|\mu^* - \bar{\mu}^*| < \lambda$ and$|\mu_{\hat{i}} - \bar{\mu}_{\hat{i}}| < \lambda$. Solving the last two inequalities gives us $\mu^* - \mu_{\hat{i}} \leq 2\lambda$. Choosing $\lambda = \epsilon/2$ we get $\mu^* - \mu_{\hat{i}} \leq \epsilon$ with probability $1 - \delta$ for $T = nK$ trails where:

$$\delta = 2Ke^{-2n\lambda^2} = 2Ke^{-\frac{T}{2K}\epsilon^2}$$

$$\implies T = \frac{2K}{\epsilon^2} \ln \frac{2K}{\delta} = \mathcal{O}(\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$$

Answer 2.

i) Let us write down the definition of the value functions according to their definition:

$$V_M^\pi(s) = \mathbb{E}[G_t|S_t = s] = \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i}|S_t = s]$$

$$V_{\bar{M}}^\pi(s) = \mathbb{E}[\bar{G}_t|S_t = s] = \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1} \bar{R}_{t+i}|S_t = s]$$

For any policy $\pi(a|s)$ we have

$$\bar{R}(s) = \sum_a \pi(a|s)\bar{R}(s, a) = \sum_a \pi(a|s)(R(s, a) + \mathcal{N}(\mu, \sigma^2))$$

$$\implies \bar{R}(s) = R(s) + \mathcal{N}(\mu, \sigma^2)$$

This gives us:

$$V_{\bar{M}}^\pi(s) = \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1}(R_{t+i} + \mathcal{N}(\mu, \sigma^2))|S_t = s]$$

$$= \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i}|S_t = s] + \mathbb{E}[\sum_{i=1}^{\infty} \gamma^{i-1}\mathcal{N}(\mu, \sigma^2)|S_t = s]$$

$$= V_M^\pi(s) + \sum_{i=1}^{\infty} \gamma^{i-1}\mathbb{E}[\mathcal{N}(\mu, \sigma^2)] = V_M^\pi(s) + \sum_{i=1}^{\infty} \gamma^{i-1}\mu$$

$$\implies V_{\bar{M}}^\pi(s) = V_M^\pi(s) + \frac{\mu}{1 - \gamma}$$

i) We are given that $\bar{P} = (\alpha * P + \beta * Q)$ where $\alpha + \beta = 1$. let the state transition matrix following policy π be P_π and $\bar{P}_\pi$ which implies $\bar{P}_\pi = (\alpha * P_\pi + \beta * Q_\pi)$. Using Bellman equation in matrix form we have:

$$V_M^\pi = R_\pi + \gamma P_\pi V_M^\pi \quad \& \quad V_{\bar{M}}^\pi = R_\pi + \gamma \bar{P}_\pi V_{\bar{M}}^\pi$$

Subtracting both we get that :

$$V_{\bar{M}}^{\pi} - V_{M}^{\pi} = \gamma \bar{P}_{\pi} V_{\bar{M}}^{\pi} - \gamma P_{\pi} V_{M}^{\pi}$$

$$\implies V_{\bar{M}}^{\pi} - \gamma \bar{P}_{\pi} V_{\bar{M}}^{\pi} = V_{M}^{\pi} - \gamma P_{\pi} V_{M}^{\pi}$$

$$\implies (I - \gamma \bar{P}_{\pi}) V_{\bar{M}}^{\pi} = (I - \gamma P_{\pi}) V_{M}^{\pi}$$

$$\implies (\alpha(I - \gamma P_{\pi}) + \beta(I - \gamma Q_{\pi})) V_{\bar{M}}^{\pi} = (I - \gamma P_{\pi}) V_{M}^{\pi}$$

Assuming $(I - \gamma P_{\pi})$ is not singular we get:

$$(\alpha I + \beta (I - \gamma P_{\pi})^{-1} (I - \gamma Q_{\pi})) V_{\bar{M}}^{\pi} = V_{M}^{\pi}$$

$$V_{\bar{M}}^{\pi} = (\alpha I + \beta (I - \gamma P_{\pi})^{-1} (I - \gamma Q_{\pi}))^{-1} V_{M}^{\pi}$$

Answer 3..We have in this question $|V^*(s) - \hat{V}(s)| \le \epsilon \quad \forall s \in S$ and $L_{\hat{V}}(s) = V^*(s) - V_{\hat{V}}(s)$ where $V_{\hat{V}}$ is the value function obtained after evaluating greedy policy with respect to $V_{\hat{V}}$. Let us write the greedy policy:

$$\hat{a} = \pi_{\hat{V}}(s) = \arg\max_{a} \sum_{s'} p(s'|s,a)[r(s,a) + \gamma \hat{V}(s')]$$

and the optimal policy:

$$a^* = \pi_{V^*}(s) = \arg\max_{a} \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V^*(s')]$$

$$V_{\hat{V}}(s) = \sum_{s'} p(s'|s,\hat{a})[r(s,\hat{a}) + \gamma V_{\hat{V}}(s')]$$

$$= \sum_{s'} p(s'|s,\hat{a})[(r(s,\hat{a}) + \gamma \hat{V}(s')) + (\gamma V_{\hat{V}}(s') - \gamma \hat{V}(s'))]$$

$$= \sum_{s'} p(s'|s,\hat{a})[r(s,\hat{a}) + \gamma \hat{V}(s')] + \gamma \sum_{s'} p(s'|s,\hat{a})[V_{\hat{V}}(s') - \hat{V}(s')]$$

$$\ge \sum_{s'} p(s'|s,a^*)[r(s,a^*) + \gamma \hat{V}(s')] + \gamma \sum_{s'} p(s'|s,\hat{a})[V_{\hat{V}}(s') - \hat{V}(s')]$$

The last statement is true as $\hat{a}$ is the greedy action with respect to value function $\hat{V}(s)$ and so any other action a^* will result in a sub-optimal value. Now $|V^*(s) - \hat{V}(s)| \le \epsilon \implies V^*(s) - \epsilon \le \hat{V}(s) \le V^*(s) + \epsilon \quad \forall s \in S$. Substituting $\hat{V}(s)$ in the above inequality we get:

$$\ge \sum_{s'} p(s'|s,a^*)[r(s,a^*) + \gamma V^*(s') - \gamma\epsilon] + \gamma \sum_{s'} p(s'|s,\hat{a})[V_{\hat{V}}(s') - V^*(s') - \epsilon]$$

$$\ge V^*(s) - 2\gamma\epsilon + \gamma \sum_{s'} p(s'|s,\hat{a})[V_{\hat{V}}(s') - V^*(s')]$$

Rearranging the terms in the above inequality we get:

$$V^*(s) - V_{\hat{V}}(s) - \gamma \sum_{s'} p(s'|s,\hat{a})[V^*(s') - V_{\hat{V}}(s')] \leq 2\gamma\epsilon$$

$$\implies L_{\hat{V}}(s) - \gamma \sum_{s'} p(s'|s,\hat{a}) L_{\hat{V}}(s') \leq 2\gamma\epsilon$$

The above is true for all s and its corresponding $\hat{a}$. To prove the rest, it is sufficient to show that the inequality holds for the peak value of $L_{\hat{V}}(s)$. Therefore, let us assume that $\bar{s}$ has the highest $L_{\hat{V}}(s)$ among all $s \in S$ so $L_{\hat{V}}(\bar{s}) \geq L_{\hat{V}}(s') \forall s' \in S$:

$$\implies \sum_{s'} p(s'|\bar{s},\hat{a}) L_{\hat{V}}(\bar{s}) \geq \sum_{s'} p(s'|\bar{s},\hat{a}) L_{\hat{V}}(s')$$

$$\implies L_{\hat{V}}(\bar{s}) \geq \sum_{s'} p(s'|\bar{s},\hat{a}) L_{\hat{V}}(s')$$

$$\implies L_{\hat{V}}(\bar{s}) - \gamma L_{\hat{V}}(\bar{s}) \leq L_{\hat{V}}(\bar{s}) - \gamma \sum_{s'} p(s'|\bar{s},\hat{a}) L_{\hat{V}}(s') \leq 2\gamma\epsilon$$

$$\implies L_{\hat{V}}(\bar{s}) \leq \frac{2\gamma\epsilon}{1-\gamma}$$

Now as the peak value of $L_{\hat{V}}(s)$ is less than $\frac{2\gamma\epsilon}{1-\gamma}$ so for all s we can write:

$$L_{\hat{V}}(s) \leq \frac{2\gamma\epsilon}{1-\gamma}$$