Comp 767 Assignment 1

Jonathan Pearce 260672004

February 5, 2020

Problem 1.
let $\bar{\mu}_i$ be the empirical average reward for arm i after $\frac{T}{K}$ pulls of the arm. By Hoeffding's inequality we have for each of the k arms,

$$\mathbb{P}[\mu_i > \bar{\mu}_i + \epsilon] \leq e^{\frac{-2T\epsilon^2}{K}}$$

Equivalently

$$\mathbb{P}[\mu_i - \bar{\mu}_i > \epsilon] \leq e^{\frac{-2T\epsilon^2}{K}}$$

Further, by symmetry we have another form of Hoeffding's inequality

$$\mathbb{P}[\bar{\mu}_i - \mu_i > \epsilon] \leq e^{\frac{-2T\epsilon^2}{K}}$$

Combining these inequalities we obtain,

$$\mathbb{P}[\mid \mu_i - \bar{\mu}_i \mid > \epsilon] \leq 2e^{\frac{-2T\epsilon^2}{K}}$$

Using the union bound and the above adaption of Hoeffding's inequality,

$$\mathbb{P}[\bigcup_{i=1}^{K} \mid \mu_i - \bar{\mu}_i \mid < \epsilon] \leq \sum_{i=1}^{K} \mathbb{P}[\mid \mu_i - \bar{\mu}_i \mid < \epsilon] \leq 2Ke^{\frac{-2T\epsilon^2}{K}}$$

Equating to δ and solving for ϵ:

$$\delta = 2Ke^{\frac{-2T\epsilon^2}{K}}$$
$$\ln(\frac{\delta}{2K}) = \frac{-2T\epsilon^2}{K}$$
$$\sqrt{\frac{K}{2T}\ln(\frac{2K}{\delta})} = \epsilon$$

Therefore for every arm i we have that $\mid \mu_i - \bar{\mu}_i \mid \leq \epsilon = \sqrt{\frac{K}{2T}\ln(\frac{2K}{\delta})}$ with probability $1 - \delta$.

Let $\hat{\mu}^*$ be the empirical average reward for the optimal arm and let $\hat{\mu}_{\hat{i}}$ be the current maximum empirical average reward. It follows,

$$\begin{aligned}\mu^* - \mu_{\hat{i}} &= \mu^* - \hat{\mu}^* + \hat{\mu}^* - \mu_{\hat{i}} \\ &\leq \mu^* - \hat{\mu}^* + \hat{\mu}_{\hat{i}} - \mu_{\hat{i}} \\ &\leq 2\sqrt{\frac{K}{2T}\ln(\frac{2K}{\delta})}\end{aligned}$$

Equating to ϵ and solving for T:

$$\epsilon = 2\sqrt{\frac{K}{2T}\ln(\frac{2K}{\delta})}$$
$$\frac{\epsilon^2}{4} = \frac{K}{2T}\ln(\frac{2K}{\delta})$$
$$T = \frac{2K}{\epsilon^2}\ln(\frac{2K}{\delta})$$

Therefore to ensure $\mu^* - \mu_{\hat{i}} \leq \epsilon$ with probability $1 - \delta$, must have T $= O(\frac{K}{\epsilon^2}\ln(\frac{K}{\delta}))$ arm pulls.

Problem 2a. From the question we have for all $a \in A$ and $s \in S$,

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

$$\Rightarrow R(s,a) = \bar{R}(s,a) - N(\mu, \sigma^2)$$

The value function for MDP M can be written as follows,

$$V_M^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]$$

$$= \mathbb{E}_\pi[R(s_{t+1}, a_{t+1}) + \gamma R(s_{t+2}, a_{t+2}) + \gamma^2 R(s_{t+3}, a_{t+3}) + ... | s_t = s]$$

$$= \mathbb{E}_\pi[R(s_{t+1}, a_{t+1}) | s_t = s] + \gamma \mathbb{E}_\pi[R(s_{t+2}, a_{t+2}) | s_t = s] + \gamma^2 \mathbb{E}_\pi[R(s_{t+3}, a_{t+3}) | s_t = s] + ...$$

$$= \sum_{k=t+1}^{\infty} \gamma^{k-t-1} \mathbb{E}_\pi[R(s_k, a_k) | s_t = s]$$

$$= \sum_{k=t+1}^{\infty} \gamma^{k-t-1} \mathbb{E}_\pi[\bar{R}(s_k, a_k) - N(\mu, \sigma^2) | s_t = s]$$

$$= \sum_{k=t+1}^{\infty} \gamma^{k-t-1} (\mathbb{E}_\pi[\bar{R}(s_k, a_k) | s_t = s] - \mathbb{E}_\pi[N(\mu, \sigma^2) | s_t = s])$$

$$= \sum_{k=t+1}^{\infty} \gamma^{k-t-1} \mathbb{E}_\pi[\bar{R}(s_k, a_k) | s_t = s] - \sum_{k=t+1}^{\infty} \gamma^{k-t-1} \mathbb{E}_\pi[N(\mu, \sigma^2) | s_t = s]$$

$$= V_{\bar{M}}^\pi(s) - \sum_{k=t+1}^{\infty} \gamma^{k-t-1} \mu$$

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

Problem 2b. Using the vector form of the bellman equation we get,

$$R = V_M^\pi(s) - \gamma P V_M^\pi(s) = (I - \gamma P) V_M^\pi(s)$$

$$R = V_{\bar{M}}^\pi(s) - \gamma \bar{P} V_{\bar{M}}^\pi(s) = (I - \gamma \bar{P}) V_{\bar{M}}^\pi(s)$$

Equating the first two equations we get,

$$(I - \gamma \bar{P}) V_{\bar{M}}^\pi(s) = (I - \gamma P) V_M^\pi(s)$$

Because P and Q are both transition matrices and $\alpha + \beta = 1$, then $\bar{P}$ is a valid transition matrix as well. Therefore the inverse of $(I - \gamma \bar{P})$ exists.

$$V_{\bar{M}}^\pi(s) = (I - \gamma \bar{P})^{-1} (I - \gamma P) V_M^\pi(s)$$

$$= (I - \gamma(\alpha P + \beta Q))^{-1} (I - \gamma P) V_M^\pi(s)$$

$$= (I - \gamma((1-\beta) P + \beta Q))^{-1} (I - \gamma P) V_M^\pi(s)$$

$$= (I - \gamma P - \gamma \beta P + \gamma \beta Q)^{-1} (I - \gamma P) V_M^\pi(s)$$

$$= (I - \gamma P + \gamma \beta (Q - P))^{-1} (I - \gamma P) V_M^\pi(s)$$

Problem 3. Let $t \in S$ be a state that maximizes the function $L_{\hat{V}}(s)$. It follows directly that $L_{\hat{V}}(t) \geq L_{\hat{V}}(t') \; \forall t' \in S$. At state t, let a be the optimal action, formally $a = \pi^*(t)$, let a' be the action taken by the greedy policy with respect to $\hat{V}$, formally $a' = \pi_{\hat{V}}(t)$. Because $\pi_{\hat{V}}(s)$ is a greedy policy then taking action a' is as good or better than taking action a:

$$R(t,a) + \gamma \sum_{t' \in S} P_{tt'}(a)\hat{V}(t') \leq R(t,a') + \gamma \sum_{t' \in S} P_{tt'}(a')\hat{V}(t')$$

From the question we have $|V^*(s) - \hat{V}(s)| \leq \epsilon \; \forall s \in S$, therefore

$$V^*(s) - \epsilon \leq \hat{V}(s) \leq V^*(s) + \epsilon$$

It follows,

$$R(t,a) + \gamma \sum_{t' \in S} P_{tt'}(a)(V^*(t) - \epsilon) \leq R(t,a') + \gamma \sum_{t' \in S} P_{tt'}(a')(V^*(t) + \epsilon)$$

Therefore,

$$R(t,a) - R(t,a') \leq 2\gamma\epsilon + \gamma \sum_{t' \in S} [P_{tt'}(a')V^*(t) - P_{tt'}(a)V^*(t)]$$

From the question statement we have $L_{\hat{V}}(s) = V^*(s) - V_{\hat{V}}(s)$. Use DS eqn. By definition of the value function we have,

$$L_{\hat{V}}(t) = [R(t,a) - \gamma \sum_{t' \in S} P_{tt'}(a)V^*(t')] - [R(t,a') - \gamma \sum_{t' \in S} P_{tt'}(a')V_{\hat{V}}(t')]$$

$$= R(t,a) - R(t,a') + \gamma \sum_{t' \in S} [P_{tt'}(a)V^*(t') - P_{tt'}(a')V_{\hat{V}}(t')]$$

Substituting for $R(t,a) - R(t,a')$ using the inequality above,

$$L_{\hat{V}}(t) \leq 2\gamma\epsilon + \gamma \sum_{t' \in S} [P_{tt'}(a')V^*(t') - P_{tt'}(a)V^*(t') + P_{tt'}(a)V^*(t') - P_{tt'}(a')V_{\hat{V}}(t')]$$

$$= 2\gamma\epsilon + \gamma \sum_{t' \in S} P_{tt'}(a')[V^*(t') - V_{\hat{V}}(t')]$$

$$= 2\gamma\epsilon + \gamma \sum_{t' \in S} P_{tt'}(a')L_{\hat{V}}(t')$$

We stated earlier $L_{\hat{V}}(t) \geq L_{\hat{V}}(t') \; \forall t' \in S$, It follows,

$$L_{\hat{V}}(t) \leq 2\gamma\epsilon + \gamma \sum_{t' \in S} P_{tt'}(a')L_{\hat{V}}(t') \leq 2\gamma\epsilon + \gamma \sum_{t' \in S} P_{tt'}(a')L_{\hat{V}}(t)$$

Rearranging,

$$L_{\hat{V}}(t) \leq \frac{2\gamma\epsilon}{1 - \gamma \sum_{t' \in S} P_{tt'}(a')} = \frac{2\gamma\epsilon}{1 - \gamma}$$