Reinforcement Learning Assignment 1

Yifan Bai(260562421), Bhavya Patwa(260964036)

February 2020

1 Multi-Arm Bandits

We have K i.i.d distributions $\{R_i\}_{i=1}^{K}$ in $[0,1]$. Total arms pulls are T so each arm is pulled $N_t(a) = \frac{T}{K}$ times for each action.
We need to guarantee that $u^* - u_{\hat{i}} \leq \epsilon, \forall \epsilon \geq 0$ with probability $1-\delta, \delta \in (0,1)$.
We have the Hoeffding's inequality: $P[E[X] > \bar{X}_t + \mu] \leq e^{-2t\mu^2}$
Thus, applying Hoeffding's inequality to K-armed bandits, we get:

$$
\begin{aligned}
P[\mu^* - u_{\hat{i}} \leq \epsilon] &= 1-\delta \\
\therefore P[\mu^* - u_{\hat{i}} > \epsilon] &= \delta \\
\therefore P[\mu^* > u_{\hat{i}}\epsilon] &= \delta \leq e^{-2N_t(a)\epsilon^2} \\
\therefore \delta &\leq e^{-2\frac{T}{K}\epsilon^2} \\
\therefore ln\delta &\leq -2\frac{T}{K}\epsilon^2 \\
\therefore -2T &\geq ln(\delta)\frac{K}{\epsilon^2} \\
\therefore T &\leq \frac{-Kln(\delta)}{2\epsilon^2}
\end{aligned}
$$

Hence Proved.

2 Markov Decision Processes

2.1 Question i

$$\mathrm{V}^{\pi}_{\mathrm{M}}(s) = E[R_{t+1} + \gamma \mathrm{V}^{\pi}(S_t + 1)|S_t = s]$$

Expanding, we get,

$$
\begin{aligned}
\mathrm{V}^{\pi}_{\mathrm{M}}(s)) &= E[\overline{R}_{t+1} + \gamma\overline{R}_{t+2} + \gamma^2\overline{R}_{t+3} + ...|S_t = s] \\
&= E[\overline{R}_{t+1} - \mathcal{N}(\mu,\, \sigma^2) + \gamma[\overline{R}_{t+2} - \mathcal{N}(\mu,\, \sigma^2)] + \gamma^2[\overline{R}_{t+2} - \mathcal{N}(\mu,\, \sigma^2)] + ...|S_t = s]
\end{aligned}
$$

Now that the Gaussian is only of constant mean μ and variance σ, we can deduce,

$$= E[\overline{R}_{t+1} + \gamma\overline{R}_{t+2} + \gamma^2\overline{R}_{t+2} + ...|S_t = s] - E[-\mathcal{N}(\mu,\, \sigma^2) - \gamma\mathcal{N}(\mu,\, \sigma^2) - \gamma^2\mathcal{N}(\mu,\, \sigma^2) - ...|S_t = s]$$

$$= \mathrm{V}^{\pi}_{\overline{(\mathrm{M})}}(s) - \mu[1 + \gamma + \gamma^2 + ...]$$

Since $\gamma \in [0,1)$, from summation of power series we have,

$$\mathrm{V}^{\pi}_{\mathrm{M}}(s) = \mathrm{V}^{\pi}_{\overline{\mathrm{M}}}(s) - \frac{\mu}{1-\gamma}$$

2.2 Question ii

$$
\begin{aligned}
\mathrm{V}^{\pi}_{\mathrm{M}}(s)) &= R^{\pi} + \gamma\mathrm{P}^{\pi}_{\overline{\mathrm{M}}}\mathrm{V}^{\pi}_{\overline{\mathrm{M}}} \\
&= R^{\pi} + \gamma(\alpha P - \alpha Q + Q)\mathrm{V}^{\pi}_{\overline{\mathrm{M}}}
\end{aligned}
$$

$$\mathrm{V}^{\pi}_{\mathrm{M}}(s)) = (1 - \gamma(\alpha P - \alpha Q + Q)^{-1} R^{\pi}$$

Then,

$$\mathrm{V}^{\pi}_{\mathrm{M}}(s)) - \gamma\mathrm{P}^{\pi}_{\mathrm{M}}\mathrm{V}^{\pi}_{\mathrm{M}} = \mathrm{V}^{\pi}_{\overline{\mathrm{M}}}(s)) - \gamma\mathrm{P}^{\pi}_{\overline{\mathrm{M}}}\mathrm{V}^{\pi}_{\overline{\mathrm{M}}}$$

$$= \mathrm{V}^{\pi}_{\mathrm{M}}(s)) - (\alpha\gamma P^{\pi} - \alpha\gamma Q^{\pi} + \gamma Q^{\pi})\mathrm{V}^{\pi}_{\mathrm{M}}$$

Therefore,

$$\mathrm{V}^{\pi}_{\mathrm{M}} = \frac{1-\gamma P}{1-\alpha\gamma P - \alpha\gamma Q + \gamma Q}\mathrm{V}^{\pi}_{\mathrm{M}}$$

3 Policy Evaluation and Improvement

We have value function $\hat{V}$ such that $|V^*(s) - \hat{V}(s)| \leq \epsilon$. Evaluating the greedy policy $\hat{V}_{\hat{V}}$ w.r.t $\hat{V}$ will give us:

$$\hat{V}_{\hat{V}}(s) = \arg\max[R_{t+1} + \gamma \sum_{s' \in S} P^{a}_{ss'}\hat{V}_{\hat{V}}(s')]$$

1. $0 < V^*(s) - \hat{V}(s) \leq \epsilon$ Taking $L_{\hat{V}}(s) = V^*(s) - \hat{V}_{\hat{V}}(s)$, we get:

$$\begin{aligned} L_{\hat{V}}(s) &= [R_{t+1} + \gamma \sum_{s' \in S} P^{a}_{ss'}V^*(s')] - \arg\max[R_{t+1} + \gamma \sum_{s' \in S} P^{a}_{ss'}\hat{V}_{\hat{V}}(s')] \\ &= \gamma \sum_{s' \in S} P^{a}_{ss'}(V^*(s') - \hat{V}_{\hat{V}}(s'))] \end{aligned}$$

The greedy policy $\hat{V}_{\hat{V}}(s) \geq \hat{V}(s), \forall s - (1)$

By assuming all P's=1, this becomes a recursive form and it can be written as:

$\leq \gamma[\epsilon + \gamma(V^*(s'') - \hat{V}_{\hat{V}}(s''))]$

(Max difference bound remains ϵ) due to (1)

$\leq \gamma[\epsilon + \gamma\epsilon + \gamma^2\epsilon + ...]$

This becomes a geometric sum and we get the result

$\leq \frac{\gamma\epsilon}{1-\gamma}$

2. $0 > V^*(s) - \hat{V}(s) \geq -\epsilon$ Taking $L_{\hat{V}}(s) = V^*(s) - \hat{V}_{\hat{V}}(s)$, we get:

$$\begin{aligned} L_{\hat{V}}(s) &= [R_{t+1} + \gamma \sum_{s' \in S} P^{a}_{ss'}V^*(s')] - \arg\max[R_{t+1} + \gamma \sum_{s' \in S} P^{a}_{ss'}\hat{V}_{\hat{V}}(s')] \\ &= \gamma \sum_{s' \in S} P^{a}_{ss'}(V^*(s') - \hat{V}_{\hat{V}}(s'))] \end{aligned}$$

The greedy policy $\hat{V}_{\hat{V}}(s) \geq \hat{V}(s), \forall s - (2)$

By assuming all P's=1, this becomes a recursive form and it can be written as:

$\geq \gamma[-\epsilon + \gamma(V^*(s'') - \hat{V}_{\hat{V}}(s''))]$

(Max difference bound remains $-\epsilon$) due to (2)

$\geq \gamma[-\epsilon - \gamma\epsilon - \gamma^2\epsilon + ...]$

This becomes a geometric sum and we get the result

$\geq \frac{-\gamma\epsilon}{1-\gamma}$

So we get:

$$\frac{-\gamma\epsilon}{1-\gamma} \leq V^*(s) - \hat{V}_{\hat{V}}(s) \leq \frac{\gamma\epsilon}{1-\gamma}$$

adding $\frac{\gamma\epsilon}{1-\gamma}$ on

4 Statement of Contribution

Both team members contributed equally to the written part of this assignment, from brainstorming ideas, formulating solutions, to reviewing solutions and typing out the document. We hereby state that all the work presented in this report is that of the authors. The link to Google Colab notebook for programming question is: `https://colab.research.google.com/drive/13nmNXWho9nYPhSY7pQ4G8kvMPX1mwFQ5#scrollTo=uqgNsjF-piAl`