Submitted by:
Akshay Singh Rana - 260963467
Harmanpreet Singh - 260962547

1 Theory

1. **Multi-Arm Bandits**
   Answer Given $\mu^* - \mu_{\hat{i}} \leq \varepsilon, \forall \varepsilon \geq 0$ with probability $1-\delta$ for $\delta \in (0,1)$.

$$\begin{aligned} & P[\mu^* - \mu_{\hat{i}} \leq \varepsilon] = 1 - \delta \\ \implies & P[\mu^* - \mu_{\hat{i}} > \varepsilon] = \delta \end{aligned}$$

   Since each reward R_i has bounded support in $[0,1]$, we can apply Hoeffding's Inequality to rewards of the bandit conditioned on selecting action. Therefore, after T/K pulls for each arm Hoeffding's inequality holds as follow:

$$\begin{aligned} P[\mu^* - \mu_{\hat{i}} > \varepsilon] &\leq e^{\frac{-2T\varepsilon^2}{K}} \\ \delta &\leq e^{\frac{-2T\varepsilon^2}{K}} \\ \ln\delta &\leq \frac{-2T\varepsilon^2}{K} \\ -T &\geq \frac{K\ln\delta}{2\varepsilon^2} \\ T &\leq \frac{-K\ln\delta}{2\varepsilon^2} \end{aligned}$$

   Therefore, T in big O notations is $O(\frac{\ln\delta}{\varepsilon^2})$.

2. **Markov Decision Process**
   Answer i. Here we explore how the values differs if the reward function is changed in

MDPs. Utilizing $\bar{R}_s^a = R_s^a + \mathcal{N}(\mu, \sigma^2)$ below,

$$
\begin{aligned}
V_{\bar{M}}^{\pi}(s) &= E_\pi[\gamma \bar{G}_t | S_s = s] \\
&= E_\pi[\bar{R}_{t+1} + \gamma \bar{R}_{t+2} + \gamma^2 \bar{R}_{t+3} + ...| S_s = s] \\
&= E_\pi[R_{t+1} + \mathcal{N}(\mu, \sigma^2) + \gamma R_{t+2} + \gamma \mathcal{N}(\mu, \sigma^2) + \gamma^2 R_{t+3} + \gamma^2 \mathcal{N}(\mu, \sigma^2) + ...| S_s = s] \\
&= E_\pi[R_{t+1} + \gamma R_{t+2} +^2 R_{t+3} + ...| S_s = s] + E_\pi[\mathcal{N}(\mu, \sigma^2)(1 + \gamma + \gamma^2 + ...)| S_s = s] \\
&= E_\pi[\gamma \bar{G}_t | S_s = s] + E_\pi[\mathcal{N}(\mu, \sigma^2)(1 + \gamma + \gamma^2 + ...)| S_s = s] \\
&= V_M^{\pi}(s) + \frac{\mu}{1 - \gamma} \\
\implies V_M^{\pi}(s) &= V_{\bar{M}}^{\pi}(s) - \frac{\mu}{1 - \gamma} \\
\implies V_M^{\pi}(s) &= V_{\bar{M}}^{\pi}(s) - \frac{\mu}{1 - \gamma}
\end{aligned}
$$

Answer ii. Here, we explore the effects if the transition matrix is changed in two MDPs. Consider bellman equation for value function in matrix form,

$$
\begin{aligned}
V_{\bar{M}}^{\pi} &= R + \gamma \bar{P} V_{\bar{M}}^{\pi} \\
V_M^{\pi} &= R + \gamma P V_M^{\pi} \\
\implies V_{\bar{M}}^{\pi} - V_M^{\pi} &= \gamma(\bar{P} V_{\bar{M}}^{\pi} - P V_M^{\pi}) \\
V_{\bar{M}}^{\pi} - \gamma \bar{P} V_{\bar{M}}^{\pi} &= V_M^{\pi} - \gamma P V_M^{\pi} \\
(I - \gamma \bar{P}) V_{\bar{M}}^{\pi} &= (I - \gamma P) V_M^{\pi} \\
V_{\bar{M}}^{\pi} &= (I - \gamma(\alpha P + \beta Q))^{-1} (I - \gamma P) V_M^{\pi}
\end{aligned}
$$

3. **Policy Evaluation and Improvement**

Answer Given V^* be the optimal value of discrete finite state MDP, and any value function $\hat{V}$ such that $|V^*(s) - \hat{V}(s)| \leq \varepsilon$. To prove: $L_{\hat{V}}(s) \leq \frac{2\gamma\varepsilon}{1-\gamma}$, where $L_{\hat{V}}(s) = V^*(s) - V_{\hat{V}}(s)$, and $V_{\hat{V}}$ is the value function obtained after evaluating the greedy policy with respect to $\hat{V}$.

Consider action a, taken from the the state with maximum loss by following the optimal policy π_*. Similarly, action b is taken from this state after evaluating the greedy policy wrt $\hat{V}$ i.e. $\pi_{\hat{V}}$. Since action b is chosen greedily wrt $\pi_{\hat{V}}$, value function of b will be greater than or equal to the value function by choosing action a. Therefore,

$$
R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \hat{V}(s') \leq R_s^b + \gamma \sum_{s' \in S} P_{ss'}^b \hat{V}(s')
$$

For $s' \in S$, its given that $|V^*(s') - \hat{V}(s')| \leq \varepsilon$

$$\implies V^*(s') - \varepsilon \leq \hat{V}(s') \leq V^*(s') + \varepsilon$$

$$\implies R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \left(V^*(s') - \varepsilon\right) \leq R_s^b + \gamma \sum_{s' \in S} P_{ss'}^b \left(V^*(s') + \varepsilon\right)$$

$$R_s^a - R_s^b \leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} \left(P_{ss'}^b V^*(s') - P_{ss'}^a V^*(s')\right)$$

Loss for these states is defined as:

$$\begin{aligned}
L_{\hat{V}}(s) &= V^*(s) - V_{\hat{V}}(s) \\
&= \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V^*(s')\right) - \left(R_s^b + \gamma \sum_{s' \in S} P_{ss'}^b V_{\hat{V}}(s')\right) \\
&= R_s^a - R_s^b + \gamma \sum_{s' \in S} \left(P_{ss'}^a V^*(s') - P_{ss'}^b V_{\hat{V}}(s')\right)
\end{aligned}$$

Now we use the inequality of $R_s^a - R_s^b$ to upper-bound loss $L_{\hat{V}}(s)$ as computed above.

$$\begin{aligned}
L_{\hat{V}}(s) &\leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} \left(P_{ss'}^b V^*(s') - P_{ss'}^a V^*(s') + P_{ss'}^a V^*(s') - P_{ss'}^b V_{\hat{V}}(s')\right) \\
L_{\hat{V}}(s) &\leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} \left(P_{ss'}^b V^*(s') - P_{ss'}^b V_{\hat{V}}(s')\right) \\
L_{\hat{V}}(s) &\leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} P_{ss'}^b \left(V^*(s') - V_{\hat{V}}(s')\right) \\
L_{\hat{V}}(s) &\leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} P_{ss'}^b L_{\hat{V}}(s')
\end{aligned}$$

Since, $L_{\hat{V}}(s') \leq L_{\hat{V}}(s) \ \forall s' \in S$, therefore,

$$L_{\hat{V}}(s) \leq 2\gamma\varepsilon + \gamma \sum_{s' \in S} P_{ss'}^b L_{\hat{V}}(s)$$

$$L_{\hat{V}}(s) - \gamma \sum_{s' \in S} P_{ss'}^b L_{\hat{V}}(s) \leq 2\gamma\varepsilon$$

$$L_{\hat{V}}(s) \left(1 - \gamma \sum_{s' \in S} P_{ss'}^b\right) \leq 2\gamma\varepsilon$$

Since $\gamma \in [0, 1)$, and $\sum P_{ss'} = 1$, because transition probabilities sum to 1, hence $(1 - \gamma) > 0$. So we divide $1 - \gamma$ to both sides.

$$L_{\hat{V}}(s)\left(1 - \gamma\right) \leq 2\gamma\varepsilon$$

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

2 Coding

1. **Explore-Exploit in Bandits A.**
   https://colab.research.google.com/drive/1-W8I5bRlFHUhZwsovDQ6IHmXQeSpHz53

2. **Dynamic Programming A.**
   Frozen Lake:
   https://colab.research.google.com/drive/1iI1GiGdu61cGSjABM2Nu23i6CCp8b0YN

   Taxi Env:
   https://colab.research.google.com/drive/1LuxGXJTFIgU3aqBVXjx6zED_AK0455D3