Question 1 (15 points)**. Multi-Arm Bandits.**

Consider the *pure exploration* setting where a learner is confronted with a *multi-arm* bandit problem defined by K arms with reward distributions $\{R_i\}_{i=1}^K$. Each R_i has bounded support in $[0,1]$ and mean μ_i. The learner explores the arms for a number of steps and has to pick the best one after T steps: $\mu^* = \max_{i=[K]} \mu_i$. Imagine the simple setting where the agent chooses each arm uniformly for a fixed number of trials T and then picks the one with the highest mean (each arm will be pulled T/K times). Derive the value of T, i.e. the number of arm pulls - needed to guarantee that $\mu^* - \mu_{\hat{i}} \leq \varepsilon, \ \forall \varepsilon \geq 0$ with probability $1-\delta, \ \delta \in (0,1)$, where $\hat{i}$ is the arm the agent chooses after T steps of interaction. T should be in big $\mathcal{O}$ notation as a function of ε and δ. *(Tip: use union bound and a concentration inequality, e.g. Hoeffding's inequality).*

Answer 1.

Let $n := \lfloor \frac{T}{K} \rfloor$ and for each $i \in \{1, ..., K\}$, let $(R_i^1, ..., R_i^n)$ be the n rewards obtained after pulling arm i. Let $\bar{\mu}_i := \frac{1}{n} \sum_{j=1}^n R_i^j$. Let $A_i := \{(R_i^1, ..., R_i^n) : |\mu_i - \bar{\mu}_i| \geq \frac{\varepsilon}{2}\}$.

By the Hoeffding's inequality [1], for each $i \in \{1, ..., K\}$:

$$P(A_i) \leq 2e^{-2n\frac{\varepsilon^2}{4}} = 2e^{-\frac{n\varepsilon^2}{2}} \tag{1}$$

By union bound and from Equation 1 above:

$$P\left(\bigcup_{i=1}^K A_i\right) \leq \sum_{i=1}^K P(A_i) \leq 2Ke^{-\frac{n\varepsilon^2}{2}} \tag{2}$$

The complement of $\bigcup_{i=1}^K A_i$ is given by:

$$C := \bigcap_{i=1}^K \{(R_i^1, ..., R_i^n) : |\mu_i - \bar{\mu}_i| < \frac{\varepsilon}{2}\}$$

and we want to ensure that $P(C) \geq 1 - \delta$. But from Equation 2:

$$P(C) = 1 - P\left(\bigcup_{i=1}^K A_i\right) \geq 1 - \sum_i^K P(A_i) \geq 1 - 2Ke^{-\frac{n\varepsilon^2}{2}}$$

If we force:

$$1 - 2Ke^{-\frac{n\varepsilon^2}{2}} \geq 1 - \delta$$

Then, we can find the minimal value n that satisfies the condition $P(C) \geq 1 - \delta$:

$$\begin{aligned} 1 - 2Ke^{-\frac{n\varepsilon^2}{2}} &\geq 1 - \delta \\ 2Ke^{-\frac{n\varepsilon^2}{2}} &\leq \delta \\ \log 2K - \frac{n\varepsilon^2}{2} &\leq \log \delta \\ \frac{n\varepsilon^2}{2} &\geq \log \frac{2K}{\delta} \\ n &\geq \frac{2}{\varepsilon^2} \log \frac{2K}{\delta} \end{aligned}$$

Now we just need to prove that all the combinations in C imply the property $\mu^* - \mu_{\hat{i}} \leq \varepsilon$.

Without loss of generality, let $\mu^* = \mu_1$. Suppose $\mu^* - \mu_j > \varepsilon$. Using the definition of C and the previous notation, we have $\bar{\mu_j} \leq \mu_j + \frac{\varepsilon}{2}$ and $\bar{\mu_1} \geq \mu_1 - \frac{\varepsilon}{2}$. Therefore:

$$\bar{\mu_j} \leq \mu_j + \frac{\varepsilon}{2} < (\mu_1 - \varepsilon) + \frac{\varepsilon}{2} = \mu_1 - \frac{\varepsilon}{2} \leq \bar{\mu_1} \tag{3}$$

This implies that $\mu_{\hat{i}} \neq \mu_j$, which completes the proof.

Hence, $n \in \mathcal{O}\left(\frac{\log(\frac{K}{\delta})}{\varepsilon^2}\right)$, which implies that $T \in \mathcal{O}\left(\frac{K \log(\frac{K}{\delta})}{\varepsilon^2}\right)$.

Question 2 (10 points)**. Markov Decision Processes.**

In this question we consider two MDPs which differ slightly. You will analyze how the value function differs for these MDPs.

1. Consider two MDPs $M = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, R, \gamma)$ and $\bar{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \bar{R}, \gamma)$, which only differ in their reward functions. We have for any $s \in \mathcal{S},\ a \in \mathcal{A},\ \bar{R}(s,a) = R(s,a) + \mathcal{N}(\mu, \sigma^2)$, where $\mathcal{N}$ is a Gaussian with constant mean μ and variance σ^2 that does not depend on s or a. For any policy π, let V_M^π denote its value function in M and $V_{\bar{M}}^\pi$ denote its value function in $\bar{M}$. For any $s \in \mathcal{S}$, express $V_M^\pi(s)$ in terms of μ and $V_{\bar{M}}^\pi(s)$.

2. In the previous part, we saw how the values differ if the reward function is changed. In this question we explore the effects if the transition matrix is changed. Consider two MDPs $M = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, R, \gamma)$ and $M = (\mathcal{S}, \mathcal{A}, \bar{\boldsymbol{P}}, R, \gamma)$ where $\bar{\boldsymbol{P}} = (\alpha \boldsymbol{P} + \beta \boldsymbol{Q})$, $\alpha \in [0,1]$, $\beta \in [0,1]$ are constants such that $\alpha + \beta = 1$, Q is another transition matrix with the same dimension as $\boldsymbol{P}$. Let V_M^π denote the value function in M and $V_{\bar{M}}^\pi$ denote the value function in $\bar{M}$. For any $s \in \mathcal{S}$, express $V_M^\pi(s)$ in terms of μ and $V_{\bar{M}}^\pi(s)$.

Answer 2.

1. The value function of a state $s \in \mathcal{S}$ under a policy π is the expected return when starting from s and following policy π thereafter:

$$\begin{aligned} V_{\bar{M}}^\pi(s) &= \mathbb{E}_\pi\left[\bar{G}_t | S_t = s\right] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \bar{R}_{t+k+1} | S_t = s\right] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \left(R_{t+k+1} + X_k\right) | S_t = s\right] \end{aligned}$$

for a family of independent and identically distributed random variables $\{X_k\}$ of Gaussian distribution with mean μ and variance σ^2. Since expectation is a linear operation, we can write:

$$\begin{aligned} V_{\bar{M}}^\pi(s) &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right] + \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k X_k | S_t = s\right] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right] + \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_\pi\left[X_k | S_t = s\right] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right] + \sum_{k=0}^{\infty} \gamma^k \mu \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right] + \frac{1}{1-\gamma}\mu \\ &= V_M^\pi(s) + \frac{\mu}{1-\gamma} \end{aligned}$$

where we have used the fact that $\gamma < 1$ to solve the geometric series and, thus, conclude that:

$$V_M^\pi(s) = V_{\bar{M}}^\pi(s) - \frac{\mu}{1-\gamma} \quad \forall s \in \mathcal{S}$$

2. We proceed as before but this time we replace the expectation by its definition:

$$\begin{aligned}
V_M^\pi(s) &= \mathbb{E}_\pi\left[G_t | S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_t + \gamma G_{t+1} | S_t = s\right] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left\{ r + \gamma \mathbb{E}_\pi\left[G_{t+1} | S_{t+1} = s'\right] \right\} \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma V_M^\pi(s') \right] \\
&= \sum_a \pi(a|s) \sum_r r \underbrace{\sum_{s'} p(s', r|s, a)}_{p(r|s,a)} + \sum_a \pi(a|s) \sum_{s'} \underbrace{\sum_r p(s', r|s, a)}_{p(s'|s,a)} \gamma V_M^\pi(s')
\end{aligned}$$

where we have used the sums to calculate the marginal distributions $p(r|s,a)$ and $p(s'|s,a)$ out of the joint distribution $p(s', r|s, a)$. Thus, we can write:

$$\begin{aligned}
V_M^\pi(s) &= \sum_a \pi(a|s) \sum_r p(r|s,a)\, r + \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\, \gamma V_M^\pi(s') \\
&= \sum_a \pi(a|s)\, \mathbb{E}\left[R(s,a)\right] + \sum_{s'} \underbrace{\sum_a \pi(a|s)\, p(s'|s,a)}_{p_\pi(s'|s)} \gamma V_M^\pi(s') \\
&= r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s)\, V_M^\pi(s') \quad \forall s \in \mathcal{S}
\end{aligned}$$

Now we can rewrite the above equation in matrix form, for a vector of states $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^{\mathrm{T}}$, where $\{s_i\}_{i=1}^n \in \mathcal{S}$ and $n = |\mathcal{S}|$, i.e. the cardinality of $\mathcal{S}$, and $s_i \neq s_j$ if $i \neq j$, that is, all the states are distinct states of $\mathcal{S}$.

$$V_M^\pi(\boldsymbol{s}) = r_\pi(\boldsymbol{s}) + \gamma \boldsymbol{P} V_M^\pi(\boldsymbol{s}) \tag{4}$$

where $\boldsymbol{P}$ is the matrix of transition probabilities from state s to s'.

We can repeat the above reasoning to $V_{\bar{M}}^\pi$ to obtain:

$$\begin{aligned}
V_{\bar{M}}^\pi(\boldsymbol{s}) &= r_\pi(\boldsymbol{s}) + \gamma \bar{\boldsymbol{P}} V_{\bar{M}}^\pi(\boldsymbol{s}) \\
&= r_\pi(\boldsymbol{s}) + \gamma \left(\alpha \boldsymbol{P} + \beta \boldsymbol{Q}\right) V_{\bar{M}}^\pi(\boldsymbol{s})
\end{aligned} \tag{5}$$

If we subtract Equation 4 from Equation 5:

$$V_M^\pi - V_{\bar{M}}^\pi = \gamma \left[\boldsymbol{P} V_M^\pi - \alpha \boldsymbol{P} V_{\bar{M}}^\pi - \beta \boldsymbol{Q} V_{\bar{M}}^\pi\right]$$

And regrouping terms we find that:

$$V_M^\pi(\boldsymbol{s}) = \left(\boldsymbol{I} - \gamma \boldsymbol{P}\right)^{-1} \left(\boldsymbol{I} - \alpha\gamma \boldsymbol{P} - \beta\gamma \boldsymbol{Q}\right) V_{\bar{M}}^\pi(\boldsymbol{s})$$

Question 3 (15 points)**. Policy Evaluation and Improvement.**

Consider a discrete, finite state MDP whose optimal value is V^*. Let $\gamma \in [0, 1)$ be the discount factor. Consider any value function $\hat{V}$ such that $\left|V^*(s) - \hat{V}(s)\right| \leq \varepsilon$ for all states $s \in \mathcal{S}$. Prove that $L_{\hat{V}}(s) \leq \frac{2\gamma\varepsilon}{1-\gamma}$ for all s, where $L_{\hat{V}}(s) = V^*(s) - V_{\hat{V}}(s)$, $V_{\hat{V}}$ is the value function obtained after evaluating the greedy policy with respect to $\hat{V}$.

Answer 3.

Translated into equations, $V_{\hat{V}}(s)$ is explicitly given by

$$V_{\hat{V}}(s) = \mathbb{E}[R_{t+1} + \gamma V_{\hat{V}}(S_{t+1})|S_t = s, A_t = \hat{A}(s)], \tag{6}$$

where we define

$$\hat{A}(s) := \arg\max_a \mathbb{E}[R_{t+1} + \gamma \hat{V}(S_{t+1})|S_t = s, A_t = a]. \tag{7}$$

We first need a simple lemma that we will use iteratively to complete the proof of the main statement afterwards.

Lemma 3.1. *Suppose that* $\left|V^*(s) - \hat{V}(s)\right| \leq \varepsilon$ *for all states* $s \in \mathcal{S}$*. Then for any* $r \in \mathcal{S}$ *and time* k*, we have*

$$V^*(r) = \mathbb{E}_{V^*}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \leq \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] + 2\gamma\epsilon \tag{8}$$

Proof.

$$\begin{aligned}
&V^*(r) - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \\
&= \mathbb{E}_{V^*}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \\
&= \mathbb{E}_{V^*}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] - \mathbb{E}_{V^*}[R_{k+1} + \gamma \hat{V}(S_{k+1})|S_k = r] \\
&+ \mathbb{E}_{V^*}[R_{k+1} + \gamma \hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma \hat{V}(S_{k+1})|S_k = r] \\
&+ \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma \hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r]
\end{aligned}$$

We will bound each of the 3 lines individually in the last formula.

For the first term:

$$\begin{aligned}
&\mathbb{E}_{V^*}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] - \mathbb{E}_{V^*}[R_{k+1} + \gamma \hat{V}(S_{k+1})|S_k = r] \\
&= \gamma \left(\mathbb{E}_{V^*}[V^*(S_{k+1})|S_k = r] - \mathbb{E}_{V^*}[\hat{V}(S_{k+1})|S_k = r]\right) \\
&= \gamma \left(\mathbb{E}_{V^*}[V^*(S_{k+1}) - \hat{V}(S_{k+1})|S_k = r]\right) \\
&\leq \gamma \left(\mathbb{E}_{V^*}[\varepsilon|S_k = r]\right) \qquad \text{(because } \left|V^*(S_{k+1}) - \hat{V}(S_{k+1})\right| \leq \varepsilon\text{)} \\
&= \gamma\varepsilon.
\end{aligned}$$

For the second term:

$$\begin{aligned}
&\mathbb{E}_{V^*}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] \\
&= \mathbb{E}_{V^*}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] - \max_{a\in\mathcal{A}}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r, A_k = a] \\
&\leq 0,
\end{aligned}$$

where the first equality is true by the definition of $V_{\hat{V}}$ given at the beginning (the greedy policy), and the inequality is obvious.

For the third term:

$$\begin{aligned}
&\mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \\
&= \gamma\left(\mathbb{E}_{V_{\hat{V}}}[\hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[V^*(S_{k+1})|S_k = r]\right) \\
&= \gamma\left(\mathbb{E}_{V_{\hat{V}}}[\hat{V}(S_{k+1}) - V^*(S_{k+1})|S_k = r]\right) \\
&\leq \gamma\left(\mathbb{E}_{V_{\hat{V}}}[\varepsilon|S_k = r]\right) \qquad \text{(because } \left|V^*(S_{k+1}) - \hat{V}(S_{k+1})\right| \leq \varepsilon\text{)} \\
&= \gamma\varepsilon.
\end{aligned}$$

Using these 3 inequalities, we obtain:

$$\begin{aligned}
&V^*(r) - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \\
&= \mathbb{E}_{V^*}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] - \mathbb{E}_{V^*}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] \\
&+ \mathbb{E}_{V^*}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] \\
&+ \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma\hat{V}(S_{k+1})|S_k = r] - \mathbb{E}_{V_{\hat{V}}}[R_{k+1} + \gamma V^*(S_{k+1})|S_k = r] \\
&\leq \gamma\varepsilon + 0 + \gamma\varepsilon \\
&= 2\gamma\varepsilon.
\end{aligned}$$

□

Now we can use Lemma 3.1 iteratively to prove the main result:

Proof.

$$
\begin{aligned}
& V^*(s) \\
&= \mathbb{E}_{V^*}[R_{t+1} + \gamma V^*(S_{t+1}) | S_t = s] \\
&\leq \mathbb{E}_{V_{\hat{V}}}[R_{t+1} + \gamma V^*(S_{t+1}) | S_t = s] + 2\gamma\varepsilon && \text{(because of Lemma 3.1)} \\
&\leq \left(\mathbb{E}_{V_{\hat{V}}}[R_{t+1} + \gamma \mathbb{E}_{V_{\hat{V}}}[R_{t+2} + \gamma V^*(S_{t+2}) | S_{t+1}] \mid S_t = s] + 2\gamma^2\varepsilon\right) + 2\gamma\varepsilon && \text{(because of Lemma 3.1)} \\
&= \mathbb{E}_{V_{\hat{V}}}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^*(S_{t+2}) \mid S_t = s] + 2\gamma^2\varepsilon + 2\gamma\varepsilon \\
&= \mathbb{E}_{V_{\hat{V}}}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^*(S_{t+2}) \mid S_t = s] + 2\gamma\varepsilon(1+\gamma) \\
&\leq \mathbb{E}_{V_{\hat{V}}}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V^*(S_{t+3}) \mid S_t = s] + 2\gamma\varepsilon(1+\gamma+\gamma^2) && \text{(similarly as above)} \\
&\leq \dots \\
&\leq \mathbb{E}_{V_{\hat{V}}}\left[\sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i} \mid S_t = s\right] + 2\gamma\varepsilon\left(\sum_{i=1}^{\infty}\gamma^{i-1}\right) && \text{(Lemma 3.1 ad infinitum)} \\
&= V_{\hat{V}}(s) + \frac{2\gamma\varepsilon}{1-\gamma},
\end{aligned}
$$

which is exactly what we wanted to prove. □

Practical Part

Graphs and analyses for the practical part can be found on the Google Colab link below:
https://colab.research.google.com/drive/1aYjxxqdKUCTuDXT9uwA-FShvJeofhJng

The codes have been uploaded to GitHub and can be found on the link:
https://github.com/mbenitah/COMP767_HW01

Statement of Contributions

For the theoretical part, before putting anything down in writing, we thoroughly discussed all three questions together. When we finally reached a consensus (specially for Question 3), Louis-François wrote Questions 1 and 3 and Michael wrote Question 2. Both of us made minor corrections or improved text comprehension for each other's part.

For the practical part, Michael did the base code of Question 1 and implemented UCB and Boltzmann. Louis-François implemented Thompson and corrected some of the logic in the implementations of UCB and Boltzmann. For Question 2, Michael implemented both algorithms as described in [2] and Louis-François worked on obtaining training and testing results.

Throughout the whole duration of the assignment, we constantly talked to each other, found some errors and commented on each other's part. All the text has been proofread by us both.

Michael Benitah
michael.benitah@umontreal.ca

Louis François Preville-Ratelle
louis-francois.preville-ratelle@umontreal.ca

References

[1] Clayton Scott. Hoeffding's inequality. http://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/notes/03_hoeffding.pdf, Winter 2014.

[2] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.