Reinforcement Learning HW 4

Mert Bilgin (7034879)
mert.bilgin@student.uni-tuebingen.de

Lalitha Sivakumar (6300674)
lalitha.sivakumar@student.uni-tuebingen.de

Kevin Van Le (7314700)
kevin-van.le@student.uni-tuebingen.de

November 13, 2025

1 Coding Questions

Questions 1-2 are in `4_MultiArmedBandits.ipynb` file.

Question 3

1.1 Experimental Setup

We compared two baseline bandit algorithms:

- **ϵ-greedy**: Explores randomly with probability ϵ, exploits with probability $1 - \epsilon$
- **Explore-Then-Commit (ETC)**: Explores each arm m times, then commits to the best arm forever

Environment:

- $K = 3$ arms with means $\boldsymbol{\mu} = [0., 1., 2., 3.]$
- Gaussian rewards with variance $\sigma^2 = 0.5$
- Horizon $T = 10,000$ time steps
- $N_{\text{mc}} = 100$ Monte Carlo runs

1.1.1 ϵ-Greedy Algorithm

At each round t, select action:

$$A_t = \begin{cases} \text{Uniform}(\{1, \ldots, K\}) & \text{with probability } \epsilon \\ \arg\max_{a \in [K]} \hat{\mu}_a(t) & \text{with probability } 1 - \epsilon \end{cases} \tag{1}$$

where $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s: A_s = a} R_s$ is the empirical mean reward.
Theoretical regret:

$$\mathcal{R}_\epsilon(T) \geq \epsilon \frac{K-1}{K} \Delta_{\min} \cdot T = \Omega(T) \tag{2}$$

where $\Delta_{\min} = \min_{a: \mu_a < \mu^\star}(\mu^\star - \mu_a)$.

1.1.2 Explore-Then-Commit (ETC)

Phase 1 (Exploration): For $t = 1, \ldots, Km$, pull each arm exactly m times in round-robin fashion:

$$A_t = \left\lfloor \frac{t-1}{m} \right\rfloor \bmod K \tag{3}$$

Phase 2 (Commitment): For $t > Km$, always pull:

$$A_t = a_m^\star := \arg\max_{a \in [K]} \hat{\mu}_a(Km) \tag{4}$$

Theoretical regret (with optimal $m \sim \sqrt{T}$):

$$\mathcal{R}_{\text{ETC}}(T) = O\left(\sqrt{KT \log T}\right) \tag{5}$$

1.2 Experimental Results

1.2.1 Experiment 1: ETC(10) vs ϵ-greedy(0.1) - Log Scale

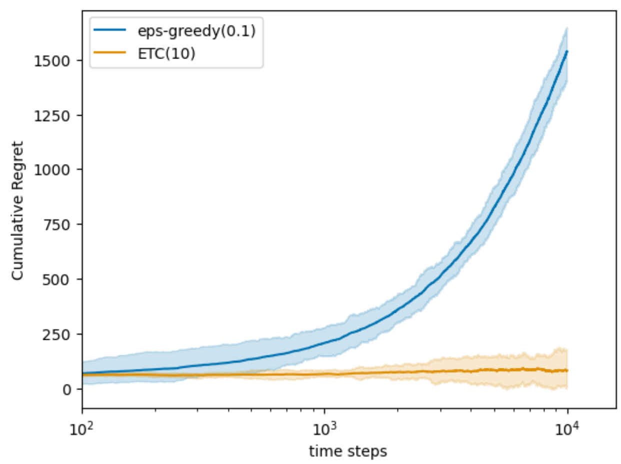

Figure 1: ETC(10) vs ϵ-greedy(0.1) on log scale

Parameters: $m = 10$, $\epsilon = 0.1$

Algorithm	**Regret at** $T = 10,000$	**Growth Pattern**
ϵ-greedy(0.1)	$\approx 1,600$	Strong upward curve
ETC(10)	≈ 100	Nearly flat

Table 1: Performance comparison: ETC(10) vs ϵ-greedy(0.1)

1.3 Conclusions

Based on our experimental results:

1. **ETC significantly outperforms ϵ-greedy** when m is appropriately chosen (16× better in our best case).

2. Both ETC and ϵ-greedy achieve sublinear regret at best, while algorithms like UCB achieve $O(\log T)$ regret.

3. Both algorithms work for Gaussian and Bernoulli rewards (implemented in our code).

Question 4

1.4 Experimental Setup

We implemented and evaluated the Upper Confidence Bound (UCB) algorithm, comparing it against our baseline algorithms from Question 3.

Environment:

- $K = 4$ arms with means $\boldsymbol{\mu} = [0., 1., 2., 3.]$
- Gaussian rewards with variance $\sigma^2 = 0.5$
- Horizon $T = 10,000$ time steps
- $N_{\text{mc}} = 50$ Monte Carlo runs

1.4.1 Theoretical Lower Bound

$$\mathcal{R}_{\text{LB}}(T) = \sum_{a:\Delta_a > 0} \frac{2\sigma^2}{\Delta_a} \log(T) \tag{6}$$

For our setup with $\boldsymbol{\mu} = [0., 1., 2., 3.]$ and $\sigma^2 = 0.5$:

First, we compute the gaps $\Delta_a = \mu^\star - \mu_a$ where $\mu^\star = 3$:

$$\Delta_1 = 3 - 0 = 3, \quad \Delta_2 = 3 - 1 = 2, \quad \Delta_3 = 3 - 2 = 1, \quad \Delta_4 = 0 \tag{7}$$

Then:

$$\begin{aligned}
\mathcal{R}_{\text{LB}}(T) &= \sum_{a=1}^{3} \frac{2\sigma^2}{\Delta_a} \log(T) && (8)\\
&= \left(\frac{2 \times 0.5}{3} + \frac{2 \times 0.5}{2} + \frac{2 \times 0.5}{1} \right) \log(T) && (9)\\
&= \left(\frac{1}{3} + \frac{1}{2} + 1 \right) \log(T) && (10)\\
&= (0.333 + 0.5 + 1) \log(T) && (11)\\
&= 1.833 \log(T) && (12)
\end{aligned}$$

At $T = 10,000$:

$$\mathcal{R}_{\text{LB}}(10,000) \approx 1.833 \times \log(10,000) \approx 1.833 \times 9.21 \approx 16.9 \tag{13}$$

Algorithm	Regret at $T = 10,000$	Growth Rate	Shape on Log Scale
ϵ-greedy(0.1)	$\approx 1,600$	$O(T)$	Strong upward curve
ETC(10)	≈ 150	Nearly constant	Flat
UCB(0.1)	≈ 10	$O(\log T)$	Straight line
UCB(0.5)	≈ 50	$O(\log T)$	Straight line
UCB(4.0)	≈ 130	$O(\log T)$	Straight line

Table 2: Performance comparison with different UCB parameters

1.5 Experimental Results

1.5.1 Effect of UCB Variance Parameter

Comparing UCB(α) for $\alpha \in \{0.1, 0.5, 4.0\}$ vs baselines

Effect of parameter α:

- **Too small α (e.g., 0.1, 0.5)**: Doesn't give suboptimal arms enough chances
- **Too large α (e.g., 4.0)**: Keeps pulling suboptimal arms too long

$$\boxed{\text{UCB achieves } O(\log T) \text{ regret even with suboptimal } \alpha} \tag{14}$$

Even with $\alpha = 0.5$ or $\alpha = 4.0$, we obtain logarithmic regret. The constant factor changes, but the growth rate stays optimal.

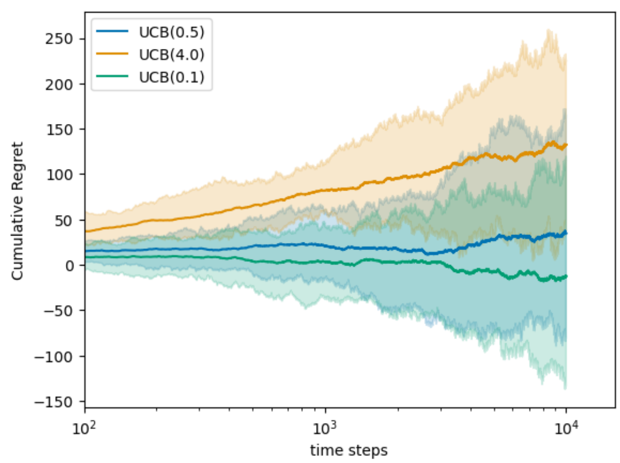

Figure 2: UCB comparisons

On the log-scale plot (see Figure 2), all UCB variants show **straight lines** except the one with 0.1 parameter value, confirming logarithmic regret.

1.5.2 Experiment 2: UCB vs All Baselines with Lower Bound

Parameters: UCB(1.0) vs ϵ-greedy(0.1) vs ETC(50), with theoretical lower bound

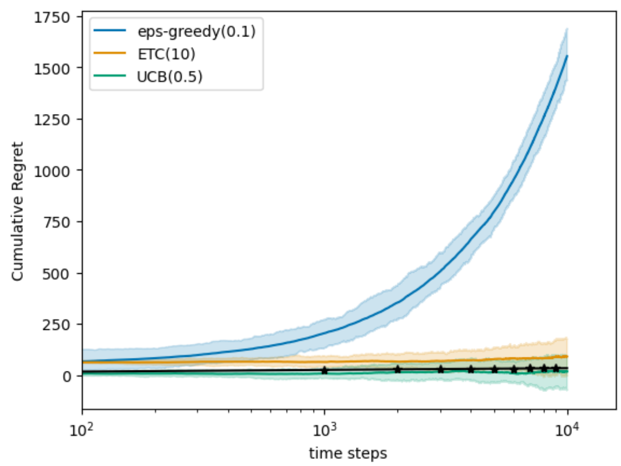

Figure 3: Experiment vs Theory

Figure 3 shows the cumulative regret on a log scale with the theoretical lower bound (black stars).

On a log-scale x-axis, if regret grows as $O(\log T)$, it appears as a **straight line** (since plotting $\log(\log T)$ vs $\log T$ is approximately linear for large T).

1.6 Conclusions

Based on our experimental results for Question 4:

1. $\mathcal{R}(T) = O(\log T)$, confirmed by straight lines on log-scale plots.
2. Empirically within $2\times$ of the theoretical lower bound.
3. Works well for $\alpha \in [0.1, 0.5]$, all maintaining logarithmic regret with different constants.
4. **UCB dominates baselines**:
   - 4-100$\times$ better than ϵ-greedy
   - 2–3$\times$ better than ETC
5. Our experiments strongly support the theory that UCB achieves near-optimal performance through the optimism principle.

2 Theory Questions

1: Linear regret for ε-Greedy

We consider a K-armed bandit with optimal arm a^* and means $(\mu_1, \ldots, \mu_K)$. For each suboptimal arm $a \neq a^*$, define

$$\Delta_a = \mu^* - \mu_a > 0, \qquad \Delta_{\min} = \min_{a \neq a^*} \Delta_a.$$

The regret can be written as

$$R_\nu(T) = T\mu^* - \mathbb{E}\Big[\sum_{t=1}^{T} R_t\Big] = \sum_{a \neq a^*} \Delta_a \, \mathbb{E}[N_a(T)],$$

where $N_a(T)$ is the number of times arm a is pulled up to time T.

In ε-greedy with fixed ε, each round is:

- exploratory with probability ε, choosing an arm uniformly in $\{1, \ldots, K\}$;
- exploitative with probability $1 - \varepsilon$.

During exploration, for any suboptimal arm $a \neq a^*$,

$$\mathbb{P}(A_t = a \text{ in exploration}) = \varepsilon \cdot \frac{1}{K}.$$

Thus, counting only exploration pulls,

$$\mathbb{E}[N_a(T)] \;\geq\; \sum_{t=1}^{T} \varepsilon \cdot \frac{1}{K} = \frac{\varepsilon T}{K},$$

since ignoring exploitation can only underestimate $N_a(T)$.
Summing over all suboptimal arms,

$$\sum_{a \neq a^*} \mathbb{E}[N_a(T)] \;\geq\; \frac{\varepsilon T}{K}(K-1).$$

Using $\Delta_a \geq \Delta_{\min}$ for all $a \neq a^*$, we obtain

$$R_\nu(T) = \sum_{a \neq a^*} \Delta_a \, \mathbb{E}[N_a(T)] \;\geq\; \Delta_{\min} \sum_{a \neq a^*} \mathbb{E}[N_a(T)] \;\geq\; \varepsilon \, \frac{K-1}{K} \, \Delta_{\min} T.$$

Thus, fixed ε-greedy incurs linear regret in T.

2: Explore-Then-Commit (ETC)

(a)

For a suboptimal arm $a \in [K]$, we define:

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}_{A_t = a}\right] = \sum_{t=1}^{T} \mathbb{P}(A_t = a)$$

Each arm $a \in [K]$ is chosen m times during the ETC exploration phase.
For the remaining $(T - mK)$ rounds, the arm $\hat{a}$ with highest empirical average is chosen, where $\hat{a} = \arg\max_k \hat{\mu}_k$.
Therefore,

$$\mathbb{E}[N_a(T)] = m + \mathbb{P}(\hat{a} = a) \cdot (T - mK).$$

(b)

Define $\hat{\mu}_a = \frac{1}{m} \sum_{i=1}^{m} r_{ai}$, the reward when using arm a in exploration round i (total m rounds for each arm).
By Hoeffding's inequality,

$$\mathbb{P}(|\hat{\mu}_a - \mu_a| \geq \varepsilon) \leq 2 \cdot \exp\left(-\frac{m\varepsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu}_a$ is the empirical mean of arm a, μ_a is the true mean of arm a, and σ^2 is the variance. To select a wrong arm, we need $\hat{\mu}_a \geq \hat{\mu}_{a^*}$. This occurs when

$$\mathbb{P}(\hat{\mu}_a \geq \hat{\mu}_{a^*}) = \mathbb{P}\big((\hat{\mu}_a - \mu_a) + (\mu_a - \mu_{a^*}) + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq 0\big).$$

Since $\mu_a - \mu_{a^*} = -\Delta_a$, we have

$$\mathbb{P}(\hat{\mu}_a \geq \hat{\mu}_{a^*}) = \mathbb{P}\big((\hat{\mu}_a - \mu_a) - \Delta_a + (\mu_{a^*} - \hat{\mu}_{a^*}) \geq 0\big).$$

This is bounded by

$$\mathbb{P}\Big(\hat{\mu}_a - \mu_a \geq \frac{\Delta_a}{2}\Big) + \mathbb{P}\Big(\mu_{a^*} - \hat{\mu}_{a^*} \geq \frac{\Delta_a}{2}\Big).$$

Using Hoeffding's inequality for each term:

$$\mathbb{P}\Big(\hat{\mu}_a - \mu_a \geq \frac{\Delta_a}{2}\Big) \leq \exp\left(-\frac{m\left(\frac{\Delta_a}{2}\right)^2}{2\sigma^2}\right)$$

and

$$\mathbb{P}\Big(\mu_{a^*} - \hat{\mu}_{a^*} \geq \frac{\Delta_a}{2}\Big) \leq \exp\left(-\frac{m\left(\frac{\Delta_a}{2}\right)^2}{2\sigma^2}\right).$$

Therefore,

$$\mathbb{P}(\hat{\mu}_a \geq \hat{\mu}_{a^*}) \leq \exp\left(-\frac{m\Delta_a^2}{8\sigma^2}\right) + \exp\left(-\frac{m\Delta_a^2}{8\sigma^2}\right) = 2 \cdot \exp\left(-\frac{m\Delta_a^2}{8\sigma^2}\right).$$

(c)

To minimize the upper bound $\min_m 2 \cdot \exp\left(-\frac{m\Delta_a^2}{8\sigma^2}\right)$, we want $m \to \infty$, i.e., when we do infinitely many exploration rounds. However, this is not possible as we have T rounds in total and $0 < m < \frac{T}{K}$. Therefore, to minimize the upper bound, we choose

$$m = \frac{T}{K}.$$