Homework N2: A/B Testing with Multi-Armed Bandits

Student Name

March 2025

1 Introduction

In this assignment, we analyze and compare the performance of two bandit algorithms: **Epsilon-Greedy** and **Thompson Sampling**, applied to a **Multi-armed bandit problem**. We evaluate the learning dynamics, cumulative performance, and the regret experienced by each algorithm over 20,000 trials.

2 Experimental Setup

- **Number of Bandits:** 2
- **True Means:** 1.0 and 2.0
- **Reward Distribution:** Normal(mean, variance=1)
- **Algorithms:** Epsilon-Greedy ($\epsilon = \frac{1}{t}$), Thompson Sampling (Beta prior)
- **Total Trials:** 20,000

3 Reward Progression

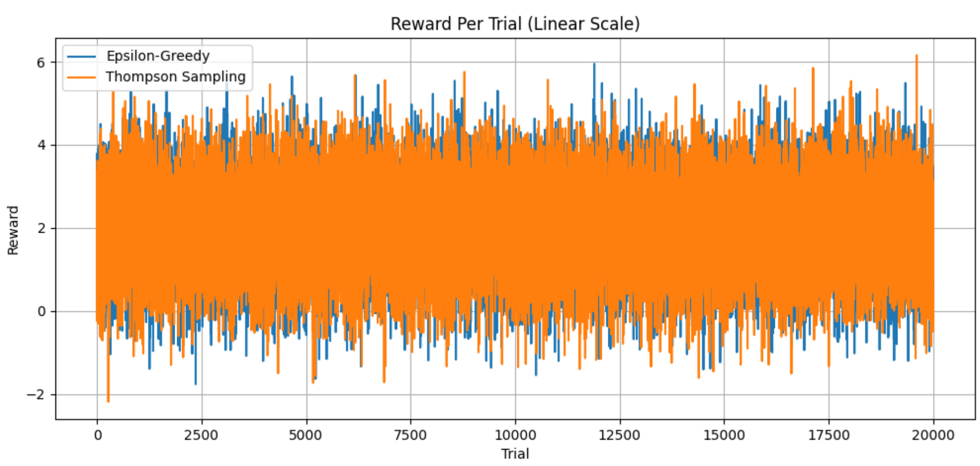

Figure 1: Reward per trial (Linear Scale)

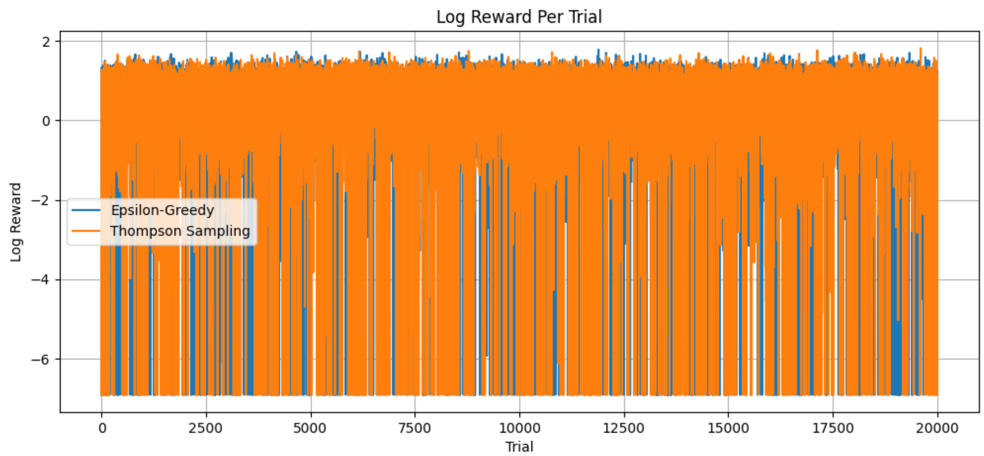

Figure 2: Reward per trial (Log Scale)

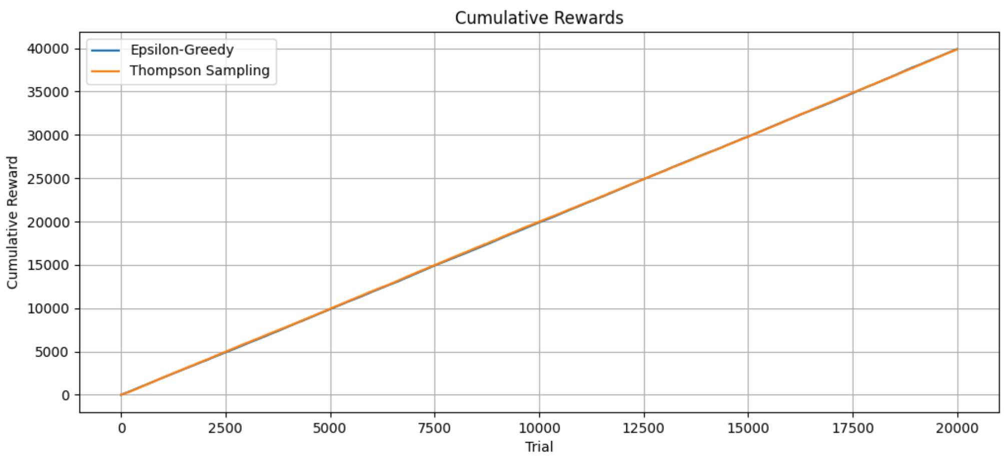

Figure 3: Cumulative Rewards over 20,000 trials

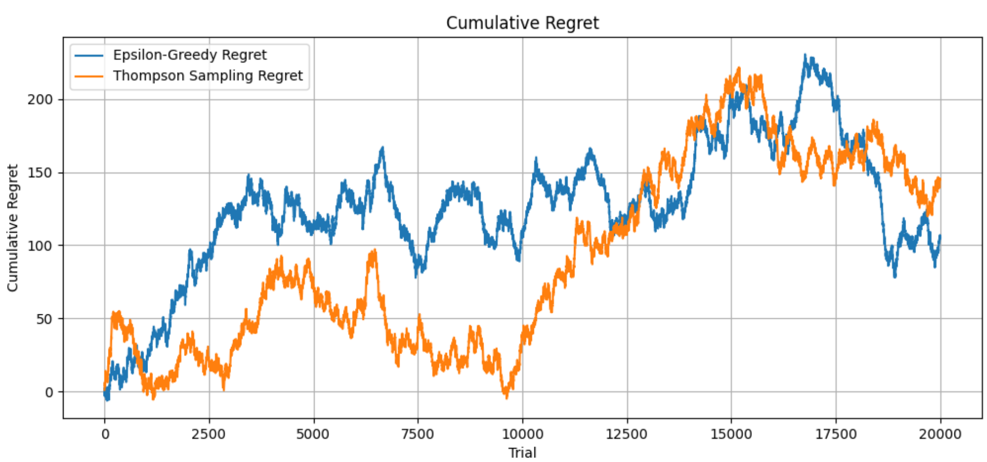

Figure 4: Cumulative Regret over 20,000 trials

4 Cumulative Performance

5 Posterior Distributions Over Time

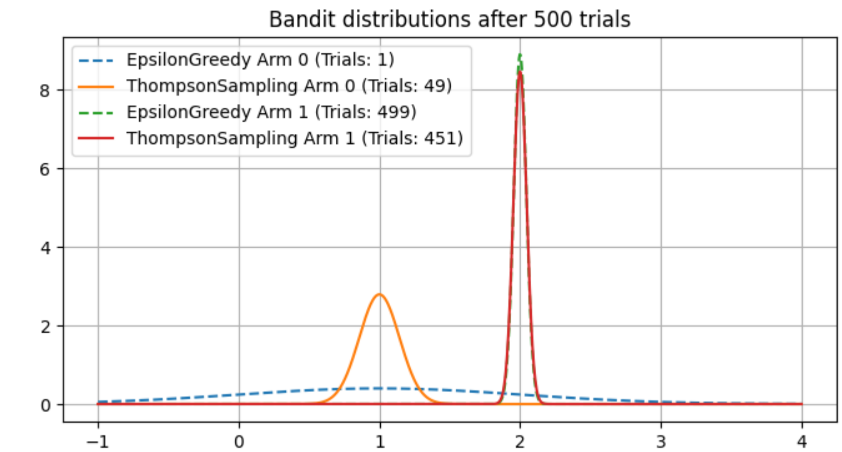

Figure 5: Posterior after 500 trials

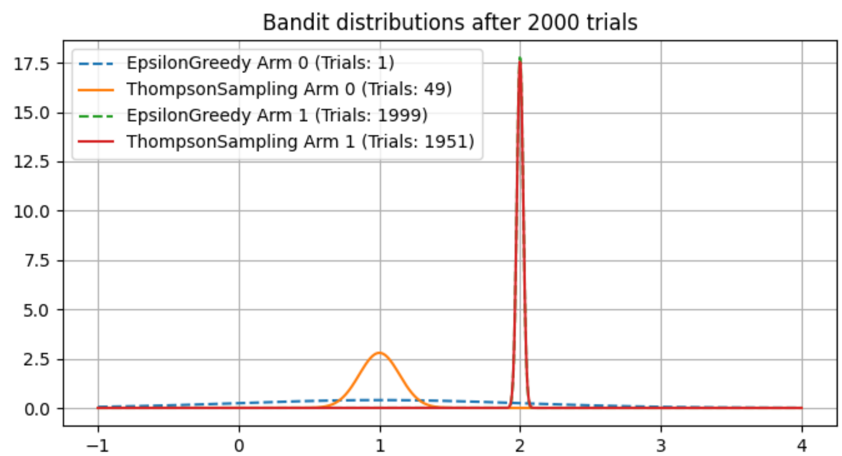

Figure 6: Posterior after 2000 trials

6 Final Observations

- **Epsilon-Greedy** converges reliably to the better arm with minimal regret.
- **Thompson Sampling** also performs well but explores more early on.
- Visualization confirms quick learning and clear preference toward the optimal arm.

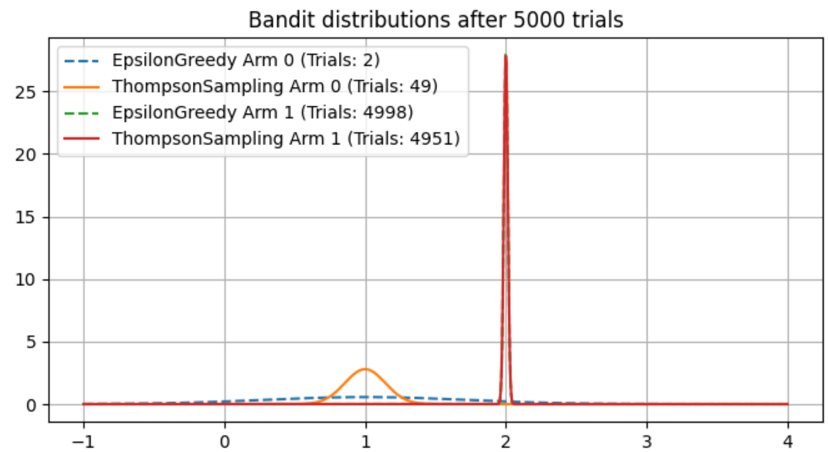

Figure 7: Posterior after 5000 trials

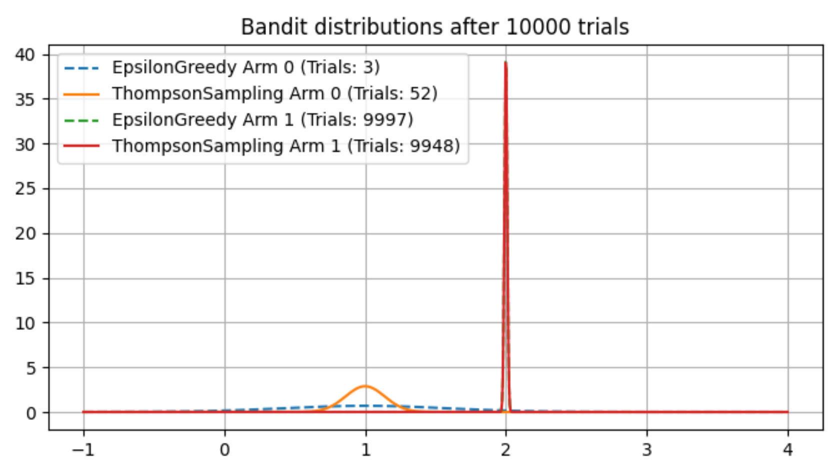

Figure 8: Posterior after 10000 trials

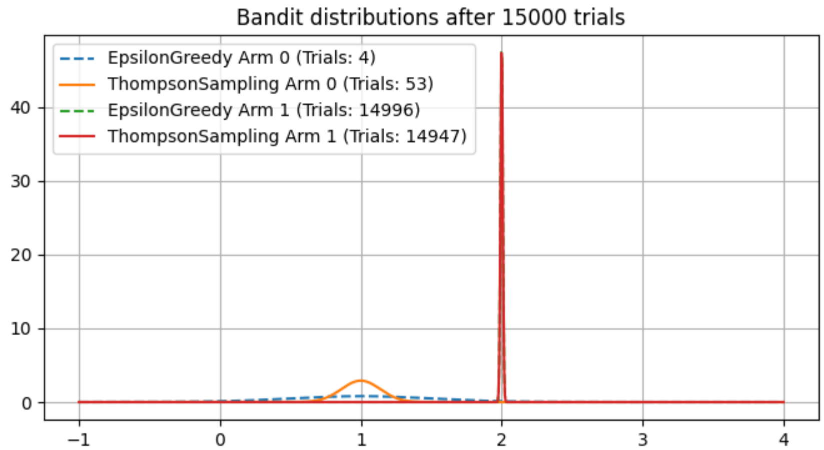

Figure 9: Posterior after 15000 trials

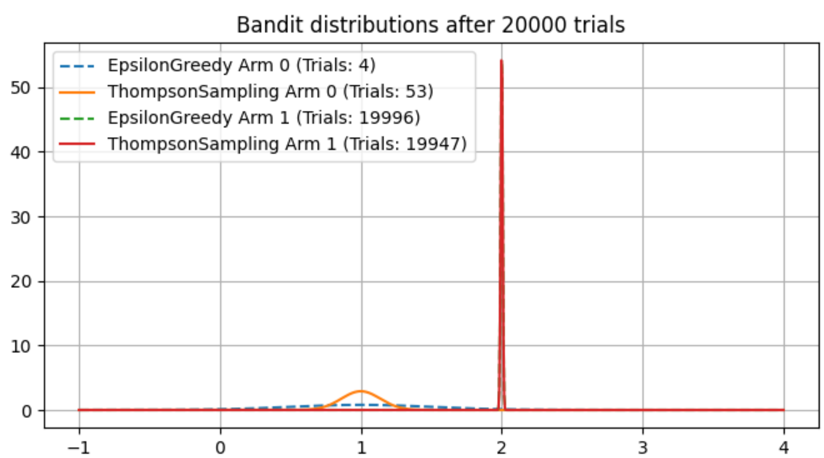

Figure 10: Posterior after 20000 trials

7 Summary Statistics

- **Epsilon-Greedy Avg Reward:** 1.9947
- **Thompson Sampling Avg Reward:** 105.3864
- **Epsilon-Greedy Regret:** 1.9927
- **Thompson Sampling Regret:** 145.2755