Comparison of ε-Greedy and UCB1 Strategies in Multi-Armed Bandits

u1541147 Tseng,Sheng Hung, u1527521 Chen,Kuan Yu

1 Problem 1: Reproducing Figure 2.1 from Sutton and Barto

Question: Your goal is to reproduce something similar to the top plot in Section 2.2, Figure 2.1 from Sutton and Barto. To make the experiment run faster, feel free to make the number of bandit tasks that you average over smaller than 2000 (but do at least 100). Discuss how your results compare with those in the Sutton and Barto book.

Answer:

Figure 1 shows the average reward over time for different ε-greedy strategies ($\varepsilon = 0, 0.01, 0.1$). The key observations are:

- $\varepsilon = 0.1$ (highest exploration) achieves the best long-term average reward.
- $\varepsilon = 0.01$ performs better than greedy ($\varepsilon = 0$) but converges more slowly.
- $\varepsilon = 0$ performs the worst since it lacks exploration and often gets stuck in suboptimal choices.

Compared to Sutton and Barto's figure, our results align well but show more variance due to averaging over fewer tasks (100 vs. 2000). Increasing the number of bandit tasks would make the results smoother.

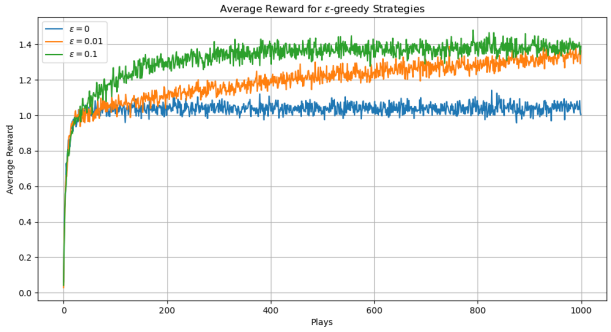

Figure 1: Average reward for ε-greedy strategies over time.

2 Problem 2: Comparing ε-Greedy with UCB1

Question: For this question, we will consider the setting of a Bernoulli bandit. This is the simplest bandit setting where each arm outputs a reward of $+1$ with probability p and a reward of 0 with probability $1 - p$. Run an experiment where you have a 10-armed Bernoulli bandit problem and compare ε-greedy ($\varepsilon = 0, 0.01, 0.1$) with UCB1. Because results will be noisy, average your results over 100 random bandit problems. To create each problem, you should sample a probability p for each arm uniformly from the range $[0, 1]$. Discuss and interpret your results.

Answer:

Figure 2 shows a comparison of different strategies:

- **Greedy** ($\varepsilon = 0$) performs the worst, as it never explores and often gets stuck in suboptimal choices.
- **Low exploration** ($\varepsilon = 0.01$) improves learning but is slow to converge.
- **Moderate exploration** ($\varepsilon = 0.1$) achieves the best balance and the highest long-term reward among ε-greedy strategies.
- **UCB1** starts off with better early exploration and outperforms $\varepsilon = 0.01$. It performs comparably to $\varepsilon = 0.1$ in the long run, but with more structured exploration.

Key Takeaways:

- UCB1 efficiently balances exploration and exploitation, outperforming $\varepsilon = 0.01$.
- $\varepsilon = 0.1$ is a simple and effective strategy, comparable to UCB1.
- Increasing the number of trials could smoothen curves and better highlight differences.

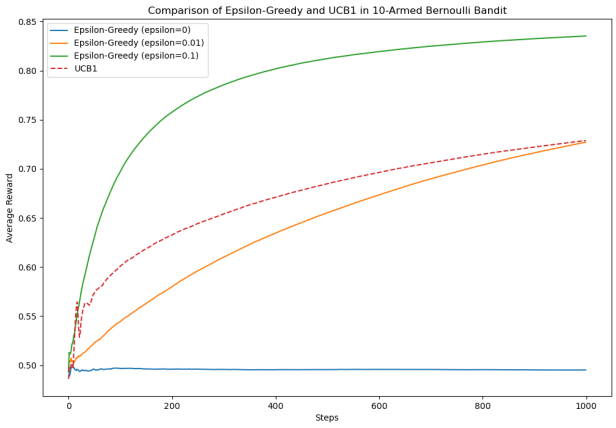

Figure 2: Comparison of ε-Greedy and UCB1 in a 10-Armed Bernoulli Bandit.