HW3: Multi-Armed Bandits

CS 6955: Adv Artificial Intelligence
University of Utah

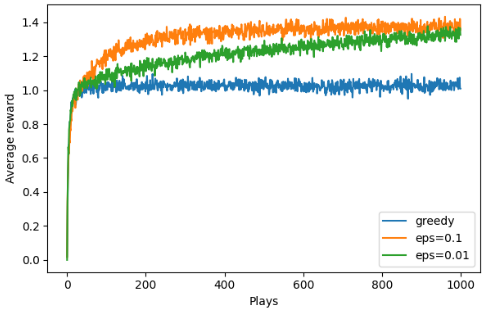
Name: Seongil Heo
UID: u1527760
Date: February 2, 2026

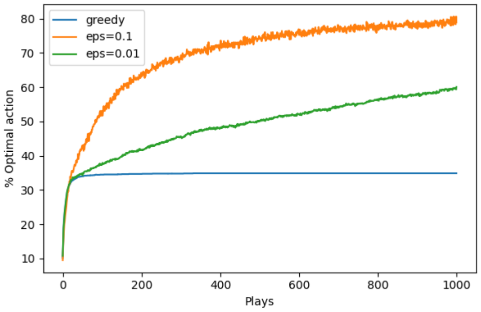
Problem 1: **Average performance of ε-greedy action-value methods on the 10-armed testbed**

Answer

The experiment used the same setting as in the book: a 10-armed bandit testbed with 2000 independent bandit tasks. The results are almost same with those reported in Sutton and Barto book. When fewer than 2000 bandit tasks are averaged, the curves appear noisier than those due to reduced averaging.

The greedy method performs poorly in the long run because it often commits early to a suboptimal action and stops exploring. In contrast, the ε-greedy methods continue to explore, increasing the probability of discovering the optimal action and resulting in higher long-term rewards. The $\varepsilon = 0.1$ method improves faster in the early stage, while ε=0.01 improves more slowly but shows stable long-term performance. These trends match the behavioral patterns described in the textbook.





Problem 2: **5-armed Bernoulli bandit**

Answer

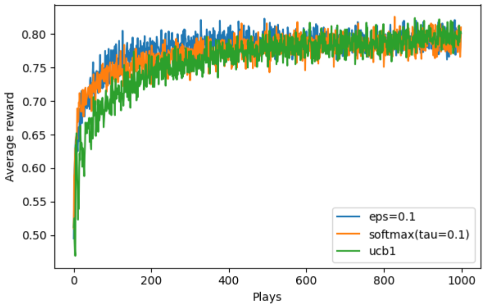
A 5-armed Bernoulli bandit was used, where each arm's success probability was sampled from a uniform distribution $U[0,1]$, rewards were binary (1 with probability p_a, 0 otherwise), and results were averaged over 1000 independent bandit problems across 1000 plays per method.

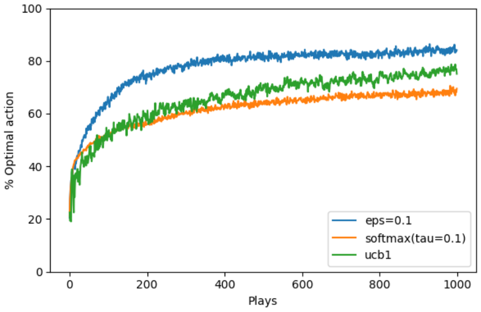
As the number of plays increased, clear differences in long-term behavior emerged among the three methods.

In the early stage, the ε-greedy method improved the fastest. This is because random exploration with a fixed probability is sufficient to quickly discover the optimal arm in a simple environment. However, as the number of plays increased, UCB1 gradually caught up and eventually achieved higher average rewards than ε-greedy. (about after 2000 plays)

This behavior can be explained by differences in exploration strategies. ε-greedy maintains a constant exploration rate, which theoretically limits the maximum probability of selecting the optimal action. In contrast, UCB1 uses an uncertainty-based bonus term, which gradually reduces exploration as uncertainty decreases. As a result, UCB1 increasingly focuses on exploiting the optimal arm over time, allowing average rewards to continue improving.

The Softmax method performs probabilistic exploration but does not explicitly account for uncertainty, leading to lower long-term performance compared to UCB1.


Average reward
Plays
eps=0.1
softmax(tau=0.1)
ucb1


% Optimal action
Plays
eps=0.1
softmax(tau=0.1)
ucb1