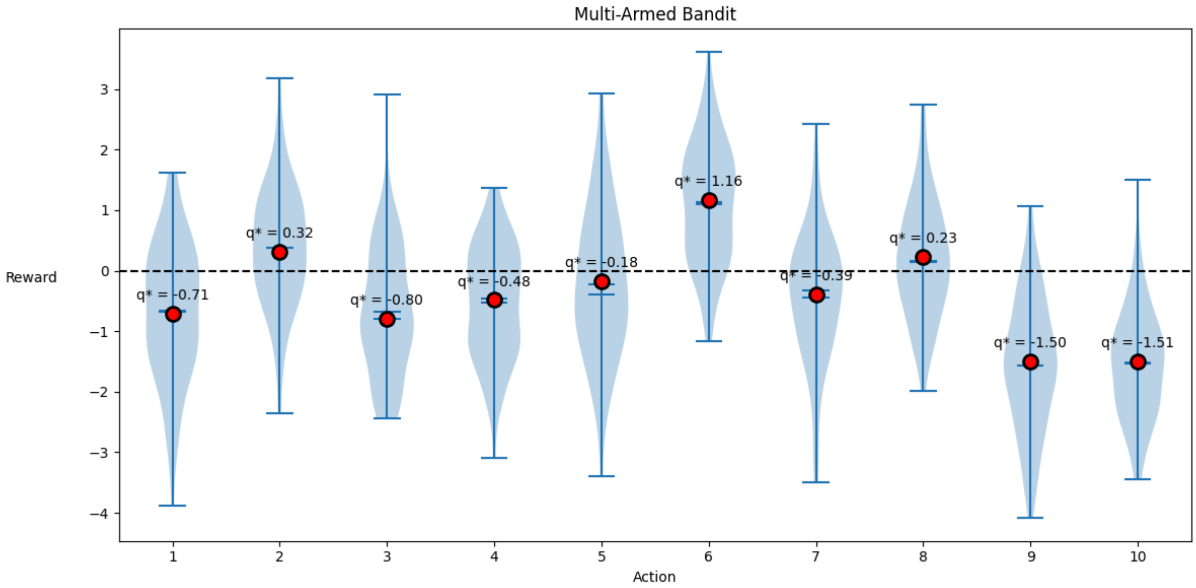
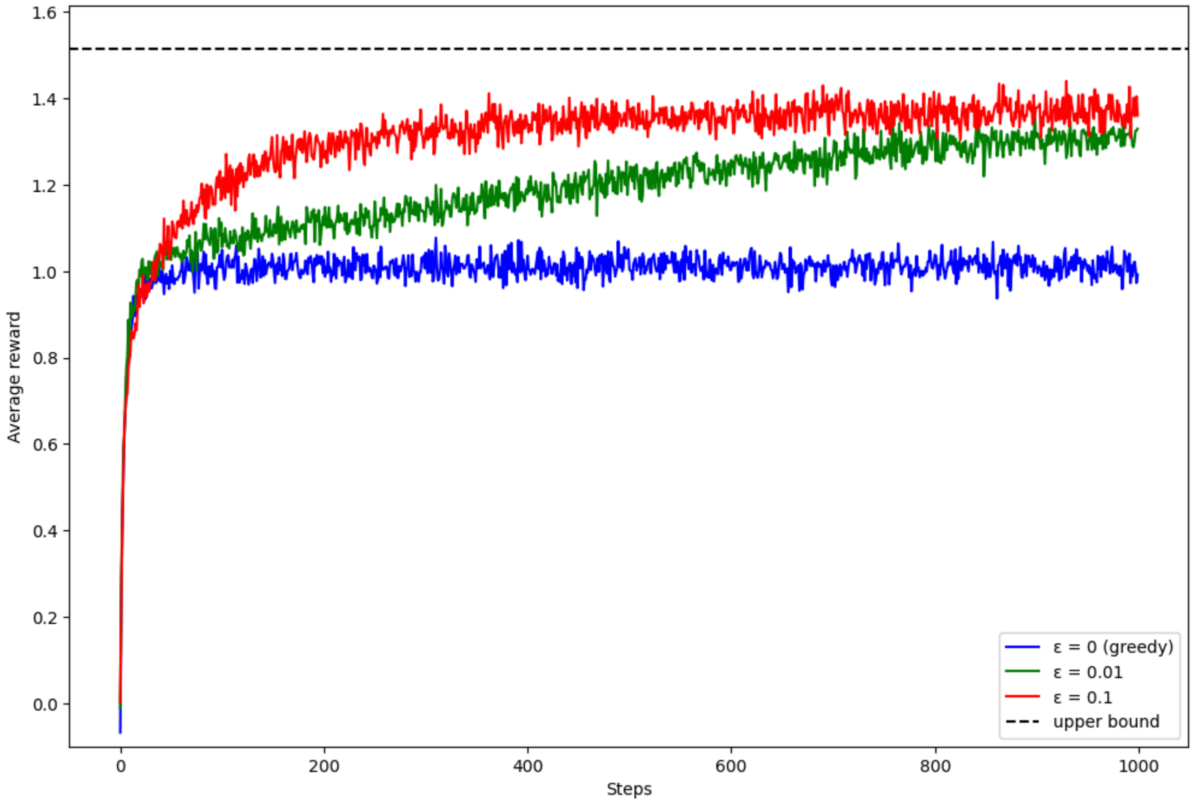
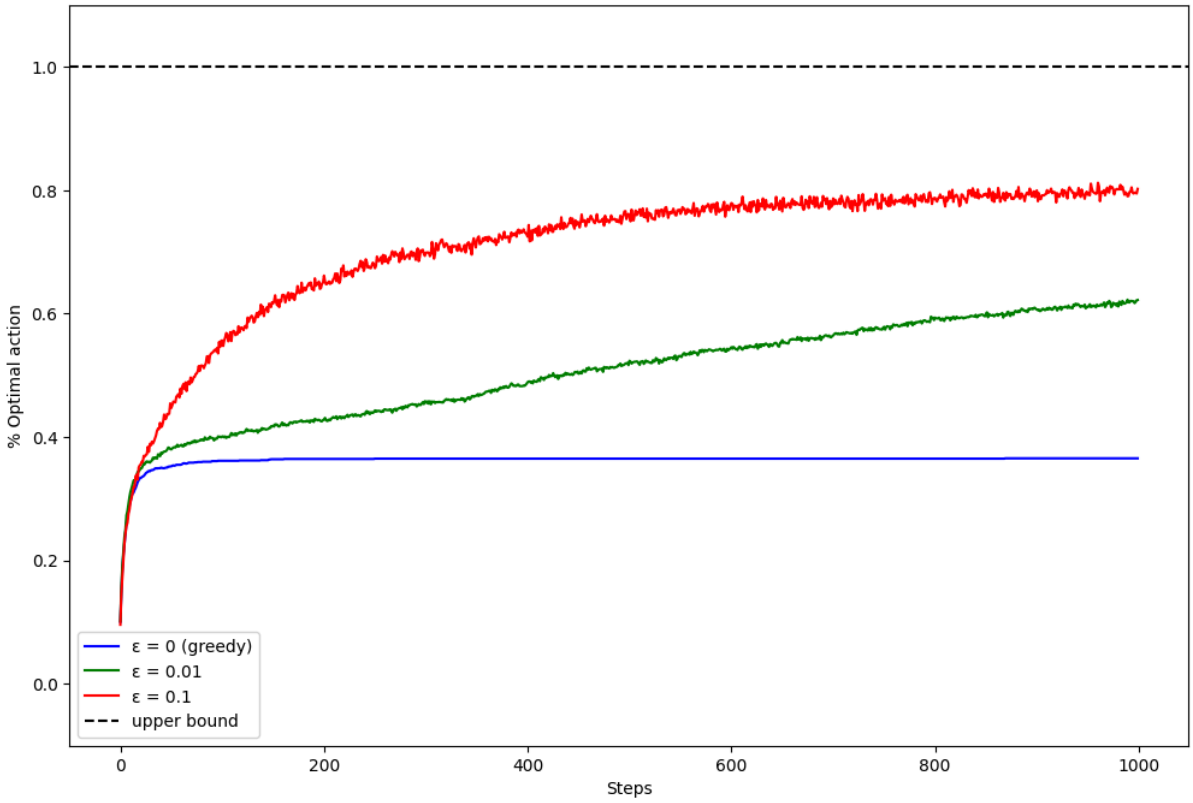
Q1- 10-armed bandit plot



Q2 — epsilon-greedy convergence

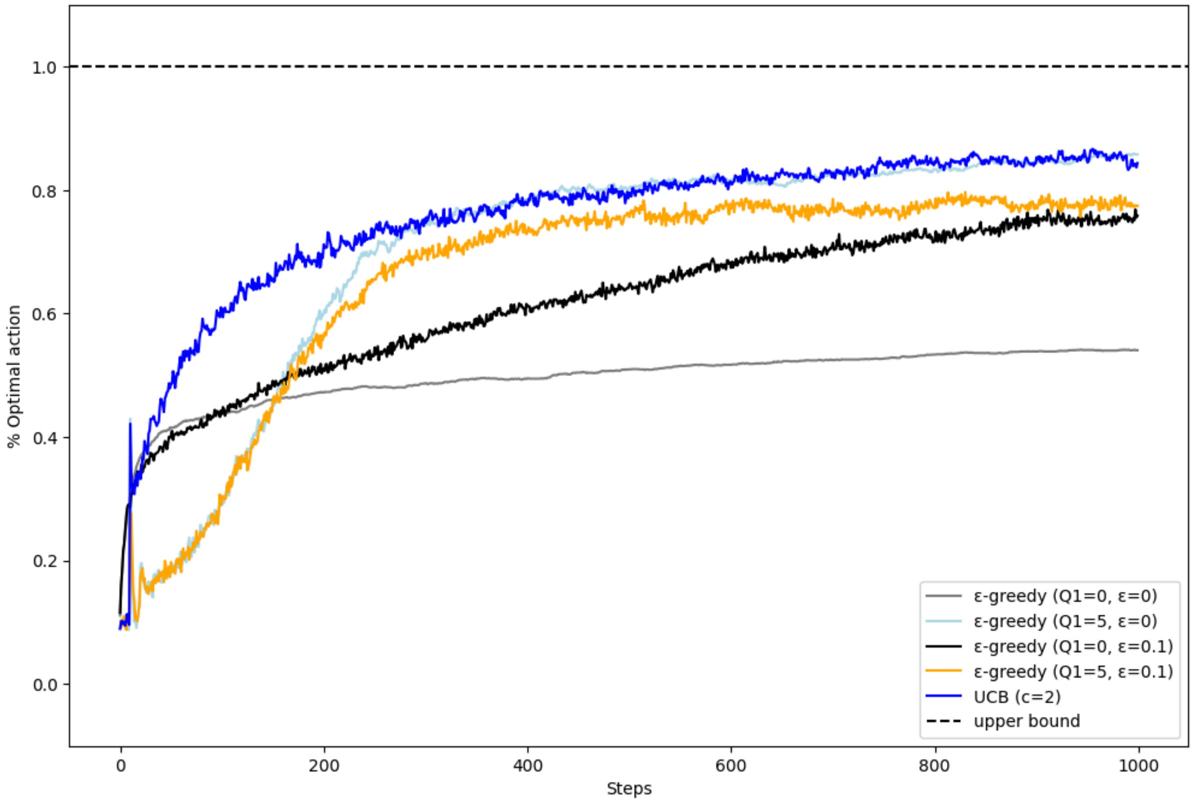
The algorithm with epsilon = 0.1 converges to the highest average reward, followed by epsilon = 0.01, while epsilon = 0 converges to the lowest reward. Larger epsilon values are explored more often, which helps the algorithm find the optimal action more reliably. Greedy methods stop exploring too early and often get stuck with suboptimal actions.

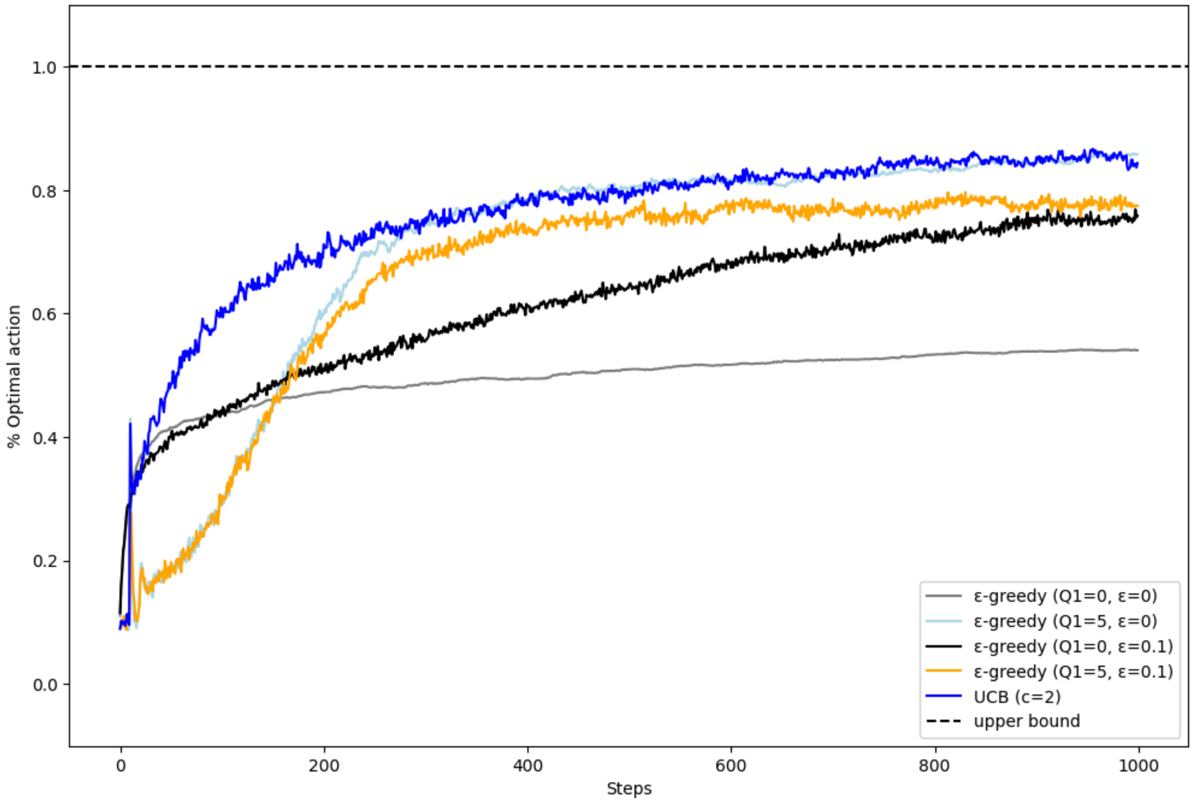




Q3 — Why spikes appear (optimistic init & UCB)

The spike occurs because the algorithm initially overestimates action values or uncertainty, which forces exploration of all actions. The sharp increase happens when high rewards are discovered early, and the decrease follows when estimates are corrected by real rewards. UCB shows a similar pattern because unvisited actions receive large exploration bonuses at the beginning.





Q4 — Weighting with non-constant step sizes

Each past reward is weighted by the product of the step sizes used when it was updated and the remaining fractions from all later updates. More recent rewards receive larger weights, while older rewards are exponentially discounted. This makes the estimate depend more on recent observations.

Q5(a) — Bias of sample-average

The sample-average estimate is unbiased because its expected value equals the true reward mean. Each reward contributes equally, and no reward is systematically over- or under-weighted.

Q5(b) — Bias with exponential average and Q1 = 0

Yes, the estimate is biased because the initial value influences all future estimates. Early rewards do not fully correct the effect of starting from zero.

Q5(c) — When Qn is unbiased

The estimate is unbiased only if the initial value equals the true expected reward. In this case, the weighted updates preserve the correct expectation.

Q5(d) — Asymptotic unbiasedness

As the number of steps increases, the influence of the initial value goes to zero. Therefore, the estimator becomes unbiased in the limit as n→infinity (n goes to infinity).

Q5(e) — Why bias is expected

Exponential recency-weighted averages prioritize recent rewards and do not fully average over all data. Because of this uneven weighting, the estimate is generally biased.