Review Paper03

1 Summary

The paper investigates reinforcement learning strategies in the context of the multi-armed bandit (MAB) problem. The authors address the classical exploration-exploitation tradeoff and analyze regret minimization strategies. Their primary contribution is proving that the optimal logarithmic regret is achievable uniformly over time for all bounded reward distributions. They introduce and analyze several index-based policies, including UCB1, UCB2, and variations of the well-known ε-greedy approach.

2 Discussion and Open Questions

- The authors assume rewards are independent across arms. Would similar regret bounds hold under weakly correlated rewards, as seen in practical settings such as recommender systems?
- The empirical evaluations suggest that tuning parameters, particularly in the ε_n-greedy method, significantly impacts performance. Could there be an adaptive method to optimize ε_n in a data-driven manner?
- Just for my curiousity, the paper's conclusion seems to be already good enough. But I am still wondering if there is a tighter lower bound for the problem. If not, it's fine.