Balancing and improving optimism and pessimism in offline-to-online learning

Group Member:Tangle; Sunjiachen
ShanghaiTech University

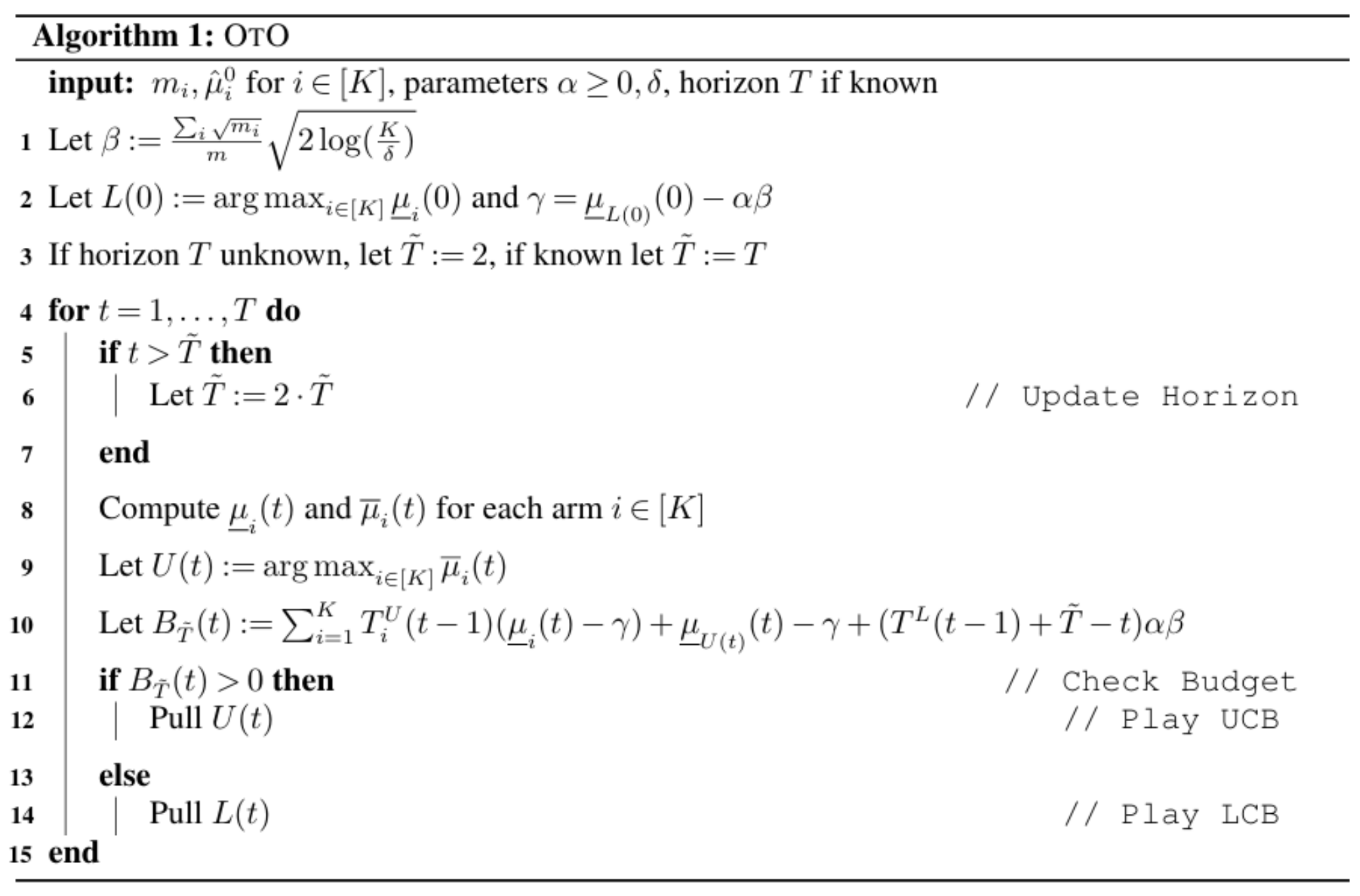
Abstract

We study the offline-to-online learning setting in stochastic finite-armed bandits, where a learner begins with offline data and then interacts with the environment to maximize cumulative reward. A key challenge lies in choosing between optimistic strategies like UCB, which perform well long-term but over-explore early, and pessimistic ones like LCB, which are better suited for short horizons. We implement and improve a novel algorithm, OTO, which can dynamically balance optimism and pessimism by computing an exploration budget. Our method adapts over time, achieving regret close to the best of UCB or LCB at any point. OTO provides strong performance guarantees and broad applicability beyond the bandit setting.

Introduction

Offline-to-online learning blends the strengths and challenges of both offline and online settings, where an agent begins with historical data and continues learning through interaction. While pessimistic strategies like LCB are effective for short horizons, and optimistic strategies like UCB excel over longer ones, neither suffices across the entire spectrum. This raises a key question: how can we balance optimism and pessimism adaptively as the learning horizon unfolds? We address this by proposing a new algorithm that dynamically adjusts its strategy, achieving low regret with respect to both the optimal policy and the logging policy. Our method is theoretically grounded and performs robustly across various offline data settings.

Algorithms



```
Algorithm 1: OTO
  input: $m_i, \hat{\mu}_i^0$ for $i \in [K]$, parameters $\alpha \geq 0, \delta$, horizon $T$ if known
Let $\beta := \frac{\sum_i \sqrt{m_i}}{m} \sqrt{2\log(\frac{K}{\delta})}$
Let $L(0) := \arg\max_{i \in [K]} \underline{\mu}_i(0)$ and $\gamma = \underline{\mu}_{L(0)}(0) - \alpha\beta$
If horizon $T$ unknown, let $\tilde{T} := 2$, if known let $\tilde{T} := T$
for $t = 1, \ldots, T$ do
    if $t > \tilde{T}$ then
        Let $\tilde{T} := 2 \cdot \tilde{T}$                                  // Update Horizon
    end
    Compute $\underline{\mu}_i(t)$ and $\overline{\mu}_i(t)$ for each arm $i \in [K]$
    Let $U(t) := \arg\max_{i \in [K]} \overline{\mu}_i(t)$
    Let $B_{\tilde{T}}(t) := \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu}_i(t) - \gamma) + \underline{\mu}_{U(t)}(t) - \gamma + (T^L(t-1) + \tilde{T} - t)\alpha\beta$
    if $B_{\tilde{T}}(t) > 0$ then                                        // Check Budget
        Pull $U(t)$                                                         // Play UCB
    else
        Pull $L(t)$                                                         // Play LCB
end
```

Figure 1

OTO Algorithm

The OTO algorithm starts by calculating an initial exploration budget parameter β and a safe reward threshold γ based on the offline data. For each round *t*, it computes the upper and lower confidence bounds for each arm and identifies the UCB arm and the LCB arm, and then checks the exploration budget. If the budget is positive, it will select the UCB arm for exploration and otherwise it would choose the LCB arm for a safer option. If the horizon is unknown, it doubles the proxy horizon T whenever the current round exceeds it. The algorithm updates the empirical means and confidence bounds after each pull and repeats this process until the end of the horizon, dynamically balancing between exploration and exploitation.

Experiments and Results

The experiment uses synthetic data to compare the performance of the OTO and SoftOTO(improved) algorithm with UCB and LCB across different scenarios. The experiment involves two instances of a multi-armed bandit problem with 20 arms and 2000 offline samples, where the first 10 arms are sampled 200 times each, and the remaining arms are not sampled at all. The reward distributions are Bernoulli with different means for optimal and suboptimal arms. The experiment evaluates the algorithms for both short (T = 200) and long (T = 2000) horizons, with the horizon known or unknown to the algorithm. The results show that OTOs effectively balances between LCB and UCB, outperforming them in different scenarios and horizon lengths.The following figure shows the experimental results.

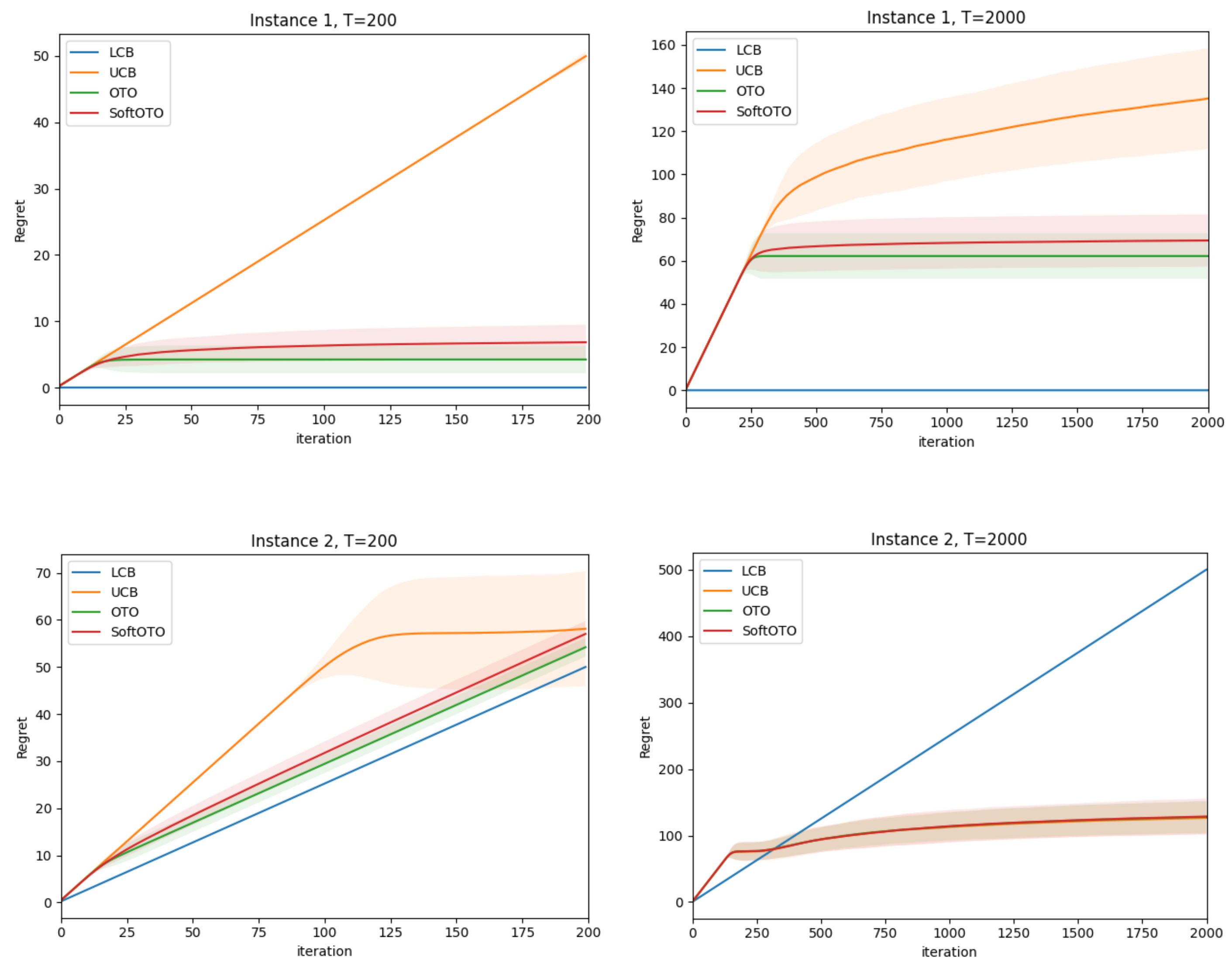


Figure 2

Experiment Results of Known Horizons

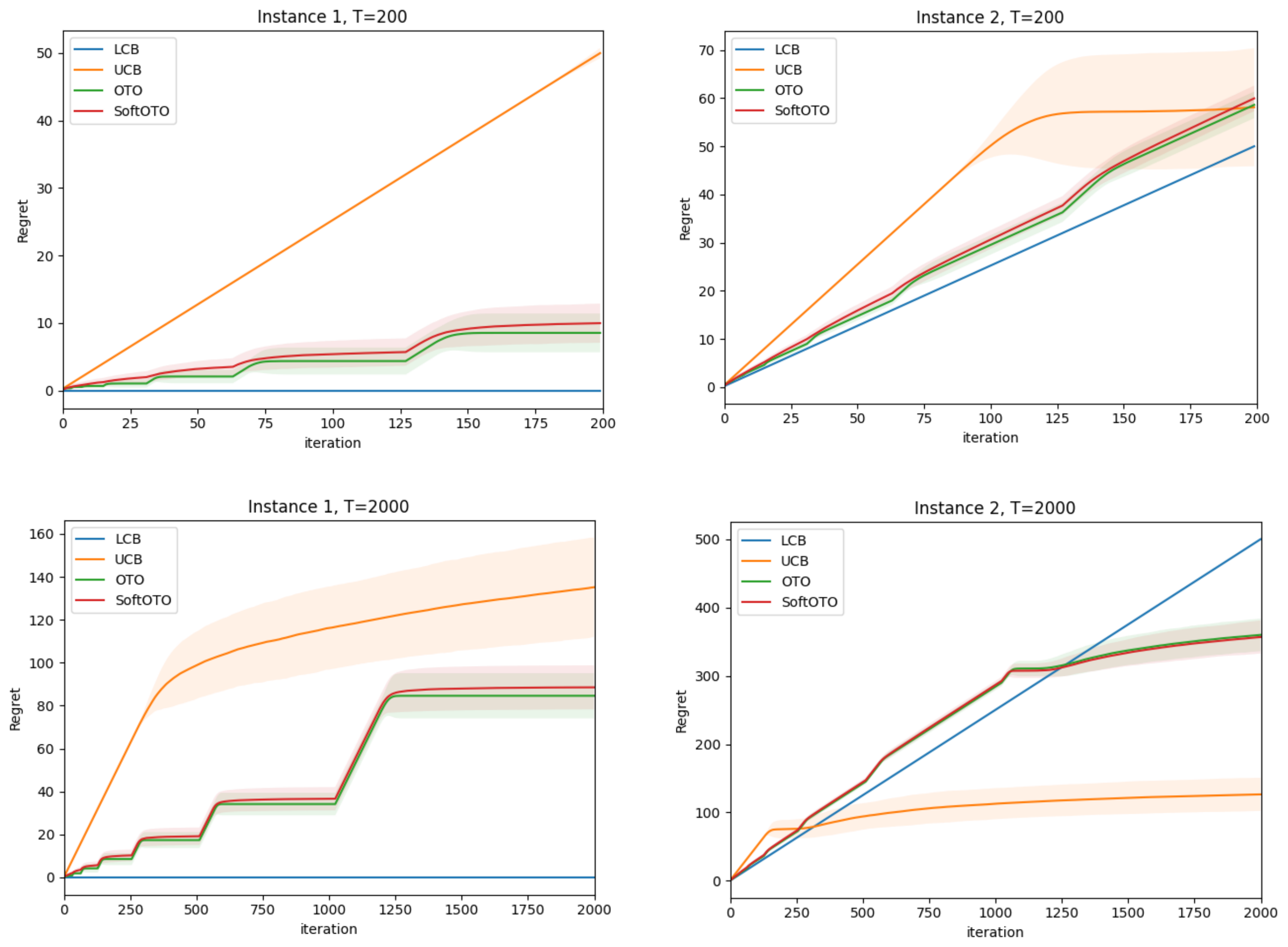


Figure 3

Experiment Results of Unknown Horizons

The results demonstrate that OTOs effectively interpolates between the strengths of LCB and UCB, achieving robust performance across different horizon lengths and problem instances. OTO adapts dynamically to the quality of the offline data and the length of the interaction horizon, outperforming both LCB and UCB in various scenarios.

Discussion

While reproducing the algorithm, we found that the parameters in the original code were not easy to modify and did not support online input. We added a new soft class to solve this problem, so that the results can be converged in time. In the implementation, we also found that the performance of the algorithm is sensitive to the choice of parameter α, which controls the trade-off between exploration and conservatism. In the case of sparse or skewed offline data distribution, it is particularly important to adjust the α value reasonably. The optimal α is actually dynamic under different behavior stages, different data distributions, and different budgets/cumulative returns. Therefore, the impact on the α setting is significant. We also adjusted α over time. This adaptive solution improves practical robustness, maintaining safety in the early stage while enabling active exploration in the later stage.

Conclusion

The Offline-to-Online (OTO) algorithm successfully addresses the challenge of balancing optimism and pessimism in the offline-to-online learning setting. By dynamically transitioning between the Lower Confidence Bound (LCB) and Upper Confidence Bound (UCB) strategies, OTO achieves robust performance across a wide range of scenarios, including both short and long interaction horizons. The experimental results on both synthetic and real-world datasets demonstrate that OTO can effectively adapt to different conditions, outperforming both LCB and UCB in various settings. Regarding future and unrealized work, in the offline-to-online process, some arms already have a large amount of offline data in the early stage and continue to obtain online data in the later stage, but the benefits are actually limited when turning to optimism exploration. On the other hand, staged freezing downgrading: reduce the exploration budget for arms that have "very high confidence" (extremely narrow confidence interval) in the early stage, and allocate more exploration to arms that have not been fully explored. This will make the algorithm converge faster and use resources more efficiently.

References

[1] Sentenac F, Lee I, Szepesvari C. Balancing optimism and pessimism in offline-to-online learning[J]. arXiv preprint arXiv:2502.08259, 2025.