Routine Bandits: Minimizing Regret on Recurring Problems

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study a variant of the multi-armed bandit problem in which a learner faces every day one of B many bandit instances, and call it a routine bandit. More specifically, at each period $h \in [1, H]$, the same bandit b^h is considered during $T > 1$ consecutive time steps, but the identity b^h is unknown to the learner. We assume all rewards distribution are Gaussian standard. Such a situation typically occurs in recommender systems when a learner may repeatedly serve the same user whose identity is unknown due to privacy issues. By combining bandit-identification tests with a KLUCB type strategy, we introduce the KLUCB for Routine Bandits (KLUCB-RB) algorithm. While independently running KLUCB algorithm at each period leads to a cumulative expected regret of $\Omega(H \log T)$ after H many periods, KLUCB-RB benefits from previous periods by aggregating observations from similar identified bandits, which yields a non-trivial scaling of $o(H \log T)$. We provide numerical illustration that confirm the benefit of KLUCB-RB.

1 Introduction

2 ...

3 Regret analysis

4 Sketch of proof

O-A: bla

A: bla

H: bla

L: bla

bla

5 Numerical experiments

6 Discussion

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Broader impact

References

A Proof of ...