Generalized Linear Contextual Bandits

Information Theory, Statistics, and Learning Course Project

Amirreza Velae, Amirabbas Afzali
Sharif University of Technology
{velaee, afzali}@sharif.edu

May 21, 2025

Abstract

We investigate the problem of generalized linear contextual bandits under adversarial conditions. This report presents an overview of online learning and bandit algorithms, followed by our novel problem setting and proposed algorithm. We analyze the regret bounds and discuss future directions for extending the framework to bandit feedback settings.

1 Introduction and Background

Online learning is a sequential decision-making framework where an agent makes decisions based on incoming data without access to future information. Key assumptions include bounded losses and a bounded decision set. The goal is to minimize regret relative to the best fixed decision in hindsight [2].

A classical example is the expert advice framework, where a learner aggregates predictions from multiple experts to minimize cumulative loss. The Weighted Majority (WM) algorithm assigns and updates expert weights based on their accuracy, achieving regret bounds scaling logarithmically with the number of experts [3].

Bandit algorithms extend online learning to scenarios with partial feedback: the learner only observes the reward for the chosen action. This creates an exploration-exploitation trade-off. Multi-Armed Bandits (MAB) formalize this, with applications in recommendation systems, clinical trials, and online advertising [1].

2 Problem Setting

We consider a finite many-armed bandit problem with an action space of K arms, where $K \gg T$ (time horizon). Each arm's reward is bounded in $[0, 1]$, and the learner aims to maximize cumulative reward over T rounds.

The adversary employs a history-dependent strategy: it observes the learner's past actions and assigns a uniform reward to all unseen arms, grouping them as a single virtual arm (see Figure 1). This induces a non-stationary environment where the available arms and their rewards evolve with the learner's history.

Formally, at time t,

$$r_t(a) = \begin{cases} r_t^{\text{seen}}(a), & a \in \{a_1, \dots, a_{t-1}\}, \\ r_t^{\text{unseen}}, & \text{otherwise.} \end{cases}$$

The available arms set is $A_t = H_t \cup \{u_t\}$, where H_t is the set of arms played until $t - 1$, and u_t is the virtual unseen arm.

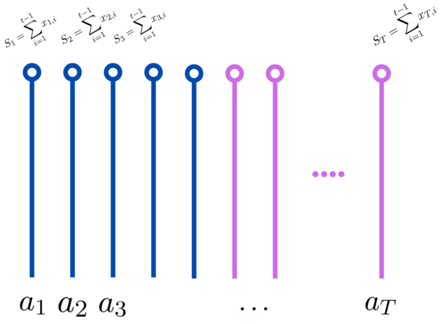
$S_1 = \sum_{i=1}^{t-1} x_{1,i}$
$S_2 = \sum_{i=1}^{t-1} x_{2,i}$
$S_3 = \sum_{i=1}^{t-1} x_{3,i}$
$S_T = \sum_{i=1}^{t-1} x_{T,i}$
....
a_1
a_2
a_3
$\dots$
a_T

Figure 1: Adaptive adversarial setting with history-dependent rewards. The learner's actions influence the reward distribution of unseen arms, creating a dynamic environment.

3 Player's Strategy

We propose a variant of the EXP3 algorithm tailored to this setting. It maintains cumulative rewards $S_{t,i}$ for known arms and the virtual arm, updating them based on observed rewards. The action probability distribution balances exploiting known arms with exploring new arms, whose probability decreases exponentially as their estimated reward accumulates.

At round t, the selection probabilities are:

$$p_{t,i} = \frac{\exp(\eta S_{t,i})}{\sum_{j=1}^{k_t} \exp(\eta S_{t,j}) + (T - k_t) \exp(\eta S_{t,k_t+1})}$$

$$p_{t,k_t+1} = \frac{(T - k_t) \exp(\eta S_{t,k_t+1})}{\sum_{j=1}^{k_t} \exp(\eta S_{t,j}) + (T - k_t) \exp(\eta S_{t,k_t+1})}.$$

for $i \in [k_t]$ and k_t is the number of arms played until round t. The parameter η controls the exploration-exploitation trade-off.

The learner samples action $A_t \sim p_t$, updates $S_{t+1,i}$ accordingly, and increments k_t if a new arm is selected.

4 Regret Bounds and Results

Our analysis shows that, under this adaptive adversarial setting, the algorithm achieves a sublinear regret bound in the online optimization regime:

$$R_T \leq \sqrt{n \log n},$$

where n is the number of rounds. Proof Sketch. The detailed proof leverages the analysis of the EXP3 algorithm. For a complete proof, please refer to the EXP3 algorithm in [1].

Note that this bound is not tight, since it hold for evry unstructured adversarial bandit with bounded rewards. The regret bound is expected to be tighter in the case of a structured adversary, where the reward distribution is more predictable.

In the bandit feedback setting, a regret bound of order $\sqrt{nk \log k}$ is expected, though establishing this rigorously remains future work.

5 Conclusion and Future Work

This work presents an online learning setting with a novel adversarial structure, where the time horizon is significantly smaller than the number of arms. We propose a modified EXP3 algorithm and establish a sublinear regret bound for online optimization.

Future work includes extending this framework to bandit feedback settings, where the learner receives partial information about the rewards of unselected arms. We also aim to explore the contextual bandit setting, where the learner receives side information about the arms (In this case, the adversary puts same reward on all unseen arms, and the learner can only observe the reward of the selected arm). This would allow us to leverage contextual information to improve the learning process.

We also aim to analyze this setting in non-stationary environments, where if the learner plays a new arm, the adversary adds a new arm to the set of arms. We argue that this setting can leverage a better regret bound, as the learner can explore new arms without being penalized by the adversary.

Acknowledgments

We thank Mr. Sarzaeem for valuable suggestions regarding the problem setting.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The Nonstochastic Multiarmed Bandit Problem, SIAM Journal on Computing, 2002.

[2] E. Hazan. Introduction to Online Convex Optimization, Cambridge University Press, 2016.

[3] N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm, Information and Computation, 1994.