Truthful Bandit Mechanisms for Repeated Two-stage Ad Auctions

Haoming Li
Shanghai Jiao Tong University
Shanghai, China
wakkkka@sjtu.edu.cn

Yumou Liu*
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
yumouliu@link.cuhk.edu.cn

Zhenzhe Zheng†
Shanghai Jiao Tong University
Shanghai, China
zhengzhenzhe@sjtu.edu.cn

Zhilin Zhang
Alibaba Group
Beijing, China
zhangzhilin.pt@alibaba-inc.com

Jian Xu
Alibaba Group
Beijing, China
xiyu.xj@alibaba-inc.com

Fan Wu
Shanghai Jiao Tong University
Shanghai, China
fwu@cs.sjtu.edu.cn

ABSTRACT

Online advertising platforms leverage a two-stage auction architecture to deliver personalized ads to users with low latency. The first stage efficiently selects a small subset of promising candidates out of the complete pool of ads. In the second stage, an auction is conducted within the subset to determine the winning ad for display, using click-through-rate predictions from the second-stage machine learning model. In this work, we investigate the online learning process of the first-stage subset selection policy, while ensuring game-theoretic properties in repeated two-stage ad auctions. Specifically, we model the problem as designing a combinatorial bandit mechanism with a general reward function, as well as additional requirements of truthfulness and individual rationality (IR). We establish an $\Omega(T)$ regret lower bound for truthful bandit mechanisms, which demonstrates the challenge of simultaneously achieving allocation efficiency and truthfulness. To circumvent this impossibility result, we introduce truthful $\alpha-$approximation oracles and evaluate the bandit mechanism through $\alpha-$approximation regret. Two mechanisms are proposed, both of which are ex-post truthful and ex-post IR. The first mechanism is an explore-then-commit mechanism with regret $O(T^{2/3})$, and the second mechanism achieves an improved $O(\log T/\Delta_\phi^2)$ regret where Δ_ϕ is a distribution-dependent gap, but requires additional assumptions on the oracles and information about the strategic bidders.

CCS CONCEPTS

• **Theory of computation** → **Algorithmic game theory and mechanism design**; **Online learning theory**; • **Information systems** → **Online advertising**.

KEYWORDS

Mechanism Design, Online Learning, Multi-Armed Bandit, Online Advertising

ACM Reference Format:
Haoming Li, Yumou Liu, Zhenzhe Zheng, Zhilin Zhang, Jian Xu, and Fan Wu. 2024. Truthful Bandit Mechanisms for Repeated Two-stage Ad Auctions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671813

1 INTRODUCTION

Modern online advertising platforms usually serve a vast number of advertisers, who participate in ad auctions to compete for ad impressions. In order to select highly relevant ads for users and to maximize social welfare, the platforms typically rank the ads by $b_i c_i$, where b_i is the bid reported by advertiser i, and c_i is the click through rate (CTR) predicted by machine learning models of the platform. However, executing complex CTR prediction models [6, 33, 34] for millions of candidate ads within a limited response time is often infeasible due to the associated high inference cost. To be able to deliver highly personalized ads to incoming users in real time, a widely adopted approach is a two-stage structure [13, 25, 32], which is also ubiquitous in large-scale online recommendation systems [7, 10, 15, 20, 29]. The first stage focuses on efficiently generating a subset of candidates that contains enough promising ads. To ensure low latency, first-stage machine learning models are lightweight and less accurate [21, 26]. The selected subset then enters the second stage, where a sophisticated CTR model provides accurate CTR predictions. Using those predictions and the submitted bids, an ad auction is conducted within the subset to determine the winning ad for display, along with the corresponding payment.

Prior works on two-stage systems has primarily focused on the performance of subset selection policy in the first stage [20, 24, 25, 32], i.e., how to efficiently select a promising subset of candidates such that the ad allocation performance of the second stage is guaranteed. Besides, as such two-stage procedures are repeatedly executed upon sequential arrivals of users, online learning of CTR in two-stage recommendation systems has also been considered [16, 31]. However, advertising systems differ from recommendation systems by the involvement of money transfer, and hence the requirement of game-theoretic properties, e.g., truthfulness and individual rationality (IR) of auction mechanisms.

*Work was done during visiting Shanghai Jiao Tong University.
†Zhenzhe Zheng is the corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671813

In this work, we address the challenge of simultaneously incorporating online learning of the subset selection policy and game-theoretic properties in repeated two-stage ad auctions. Concretely, we consider a repeated auction with n advertisers and T rounds, where, in each round, at most k advertisers are selected to enter the second stage. Advertisers report their bids before the first round starts, to optimize their cumulative utility over T rounds. In the first stage of round t , we select a subset of advertisers denoted as K_t to enter the second stage. Then for each advertiser i inside K_t, its second-stage CTR prediction c_{it}, which we assume to be an i.i.d. sample from a probability distribution D_i, is observed. We fix the second stage to be a second-price auction, a well-known truthful mechanism prevalent in ad auctions. The second-price auction only involves advertisers within K_t, and determines the advertiser with the highest $c_{it}b_i$ to be displayed. Our goal is to maximize the cumulative social welfare without knowing the second-stage CTR distributions $D_1, \cdots, D_n$ beforehand, while preserving the truthful and IR properties of the T-round mechanism.

If we ignore the strategic behaviours of advertisers, this problem is equivalent to a combinatorial bandit with a general reward function [5], where advertisers are treated as base arms, and in each round we choose a super arm K_t subject to the cardinality constraint $|K_t| \leq k$, and receive reward $\max_{i \in K_t}\{b_i c_{it}\}$ [1]. However, the additional requirement of truthfulness makes our problem a non-trivial extension of the original bandit problem. In fact, advertisers could misreport their values to manipulate the outcome of both stages in a round. The observed samples of one round will further influence the bandit algorithm's behaviour in subsequent rounds. To reveal the challenge of simultaneously ensuring performance and truthfulness, we present in Proposition 1 the impossibility to design a stochastically truthful (please refer to Definition 5) mechanism even if the CTR distributions are known beforehand. Building upon this result, we establish an $\Omega(T)$ regret lower bound on bandit mechanisms that are stochastically truthful (Theorem 1).

To circumvent the impossibility result, we introduce truthful approximation oracles, which are constant-factor approximation algorithms for solving the offline optimization problem in a truthful manner. These oracles allow us to preserve truthfulness at the cost of sacrificing allocation efficiency. We use the notion of $\alpha-$approximation regret to evaluate bandit algorithms which calls an $\alpha-$approximation oracle. We propose a bandit mechanism with $O(T^{2/3})$ regret and achieves ex-post truthfulness and ex-post IR. The algorithm is based on a straightforward explore-then-commit (ETC) strategy. The separation of exploration phase and exploitation phase prevents strategic bidders from influencing the data collection process, thereby ensuring truthfulness. Furthermore, we discover that truthful approximation oracles often exhibit a typical structure: scoring the arms by their CTR distributions, then select the top-k arms according to the product of their bid and score. By exploiting this structure and making additional assumptions on prior knowledge of bidders' private values, we propose an algorithm that achieves an improved regret bound of $O(\log T/\Delta_\phi^2)$, while preserving ex-post truthfulness and ex-post IR, where Δ_ϕ is a distribution-dependent gap.

To summarize, our major contributions in this work include:

- To the best of our knowledge, this is the first work that jointly considers online subset selection and game-theoretic properties in the setting of repeated two-stage auctions.
- We demonstrate the difficulty of the problem by establishing a $\Omega(T)$ regret lower bound for truthful bandit mechanisms.
- We introduce truthful $\alpha-$approximation oracles, which allow us to design two ex-post truthful and ex-post IR bandit mechanisms: one has $O(T^{2/3})$ α-approximation regret, and the other one enjoys a better $O(\log T/\Delta_\phi^2)$ α-approximation regret but requires additional assumptions on both the oracle and the bidders.
- We validate the efficiency and truthfulness of our proposed mechanisms through experiments on both synthetic and real-world data, with results aligning well with our theoretical claims.

2 MODEL AND PRELIMINARIES

We consider a repeated single-slot ad auction setting with n advertisers [2] $[n]$ and T rounds, where a two-stage auction is conducted in each round. [3] The advertisers have their private values $\mathbf{v} = (v_1, \cdots, v_n)$ which is unknown to the auctioneer. Before the first round starts, all advertisers submit their bids $\mathbf{b} = (b_1, \cdots, b_n)$. We assume that all values and bids are bounded in $[0, V]$, where $V > 0$. The second-stage CTR prediction of each advertiser i follows distribution D_i, where D_i is a probability distribution over $[0, 1]$. We use $D = (D_1, \cdots, D_n)$ to denote the product distribution of each D_i. To characterize the two-stage ad auction in each round, we first define second-price auction within a subset.

Definition 1 (Second-price auction within a subset). *Given n advertisers, a subset $K \subseteq [n]$ with cardinality constraint $|K| \leq k$, CTR $\{c_i\}_{i\in K}$, and a bid vector $\mathbf{b} \in [0, V]^n$, a second-price auction within K determines the following allocation x_i and payment p_i for each $i \in [n]$:*

$$x_i = \begin{cases} 1, & \text{if } i \in K \text{ and } i = \operatorname{argmax}_{j\in K}\{b_j c_j\} \\ 0, & \text{otherwise,} \end{cases}$$

$$p_i = \begin{cases} \frac{\max_{j\in K, j\neq i} c_j b_j}{c_i}, & \text{if } i \in K \text{ and } i = \operatorname{argmax}_{j\in K}\{b_j c_j\} \\ 0, & \text{otherwise,} \end{cases}$$

where ties are broken consistently. The output of the auction is an allocation vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and a payment vector $\mathbf{p} = (p_1, p_2, \cdots, p_n)$. Throughout the paper, we use $\mathbf{x}(\mathbf{b}, \{c_i\}_{i\in K}, K)$ and $\mathbf{p}(\mathbf{b}, \{c_i\}_{i\in K}, K)$ to denote outcomes of second-price auctions [4].

In each round $t \in [T]$, a two-stage auction is conducted as follows:

[1] This reward function is "general" in the sense that the expected value $\mathbb{E}_{c_{it}\sim D_i}[\max_{i\in K_t}\{b_i c_{it}\}]$ not depend only on the means of random variables , i.e., $\mathbb{E}_{D_i}[c_{it}]$, but on the entire distributions of these variables.

[2] We use the terms advertiser, bidder, and arm interchangeably.

[3] For simplicity, we assume that the time horizon T is known beforehand, but our results can be extended to the case with unknown T using a standard "doubling trick" [2, 19].

[4] Sometimes we might slightly abuse notations and write $\mathbf{x}(\mathbf{b}, \mathbf{c}, K)$ and $\mathbf{p}(\mathbf{b}, \mathbf{c}, K)$, where $\mathbf{c} = (c_1, \cdots, c_n)$ is the full CTR vector, although $\mathbf{x}$ and $\mathbf{p}$ only depends on $\{c_i\}_{i\in K}$.

- First stage. The auctioneer selects a subset of advertisers K_t with cardinality constraint $|K_t| \leq k$ based on the bid vector $\mathbf{b}$ and all observations from previous rounds. Only the advertisers in K_t enters the second stage.
- Second stage. For each $i \in K_t$, the auctioneer observes the second-stage CTR prediction c_{it}, which is an i.i.d. sample from distribution D_i. Then a second-price auction is run within K_t (see Definition 1), using the predicted CTRs $\{c_{it}\}_{i\in K_t}$ and the submitted bid vector $\mathbf{b}$. The allocation to advertisers in this round is $\mathbf{x}_t = \mathbf{x}(\mathbf{b}, \{c_{it}\}_{i\in K_t}, K_t)$, with payment $\mathbf{p}_t = \mathbf{p}(\mathbf{b}, \{c_{it}\}_{i\in K_t}, K_t)$.

Our goal is to maximize the cumulative social welfare of T rounds. The reward we obtain in each round t is defined as the social welfare of the second-price auction in that round, i.e., $R_t = \max_{i\in K_t}\{b_i c_{it}\}$. Since all bids are in $[0, V]$ and CTRs are in $[0, 1]$, the reward R_t is in $[0, V]$. The reward depends on sampled CTR predictions, and we define its expectation with respect to distribution D as $R_D(K_t) = \mathbb{E}_D[R_t]$. Then $R_D(K_t)$ is a scalar that depends on K_t, D, and $\mathbf{b}$, serving as a measure of how good K_t is, given D and $\mathbf{b}$.

We define the offline optimal reward, i.e. the maximum expected reward one can achieve if D (and $\mathbf{b}$) is known, as $\mathrm{OPT}_D = \max_K R_D(K)$.

To maximize the cumulative social welfare, we should select proper advertisers to enter the second stage in each round. Moreover, we can only observe the second-stage CTR predictions of advertisers that enters the second stage. This sequential subset selection procedure with partial feedback can be viewed as a combinatorial (semi-)bandit, where we treat advertisers as base arms and subsets as super arms. Following the convention of bandit algorithms, we define regret as the performance measure.

Definition 2. *The regret of an algorithm is*

$$Reg(T) = T \cdot \mathrm{OPT}_D - \mathbb{E}\left[\sum_{t=1}^{T} R_D(K_t)\right].$$

In our setting, a bandit algorithm also defines a T-round mechanism [5], where we consider the total allocation , $\mathbf{X} = (X_1, \cdots, X_n) = \sum_{t=1}^{T} \mathbf{x}_t$, and the total payment, $\mathbf{P} = (P_1, \cdots, P_n) = \sum_{t=1}^{T} \mathbf{p}_t$. We refer to those mechanisms as bandit mechanisms. When a fixed bandit algorithm runs on a fixed instance $(D, \mathbf{b})$ for several times, the resulting allocation $\mathbf{X}$ and payment $\mathbf{P}$ may be different, because the CTR predictions in each round are stochastic. We can consider this T-round interaction process as first drawing CTR predictions c_{it} for all $i \in [n]$ and $t \in [T]$ from the distributions D, and then running the bandit algorithm on these fixed samples. Specifically, a $n \times T$ realization table $\mathbf{C}$ whose (i, t)-th entry is the c_{it} to reveal when arm i is played in the t-th round. When a table $\mathbf{C}$ is fixed, there is no stochasticity in the mechanism, so the total allocation $\mathbf{X}(\mathbf{C}, \mathbf{b})$ and payment $\mathbf{P}(\mathbf{C}, \mathbf{b})$ of a mechanism are also fixed. This allows us to consider the utility function of advertiser i with respect to any table $\mathbf{C}$, which is a deterministic function.

[5]Since the second-stage auction mechanism is fixed to be second-price (See Definition 1), the allocation rule of a T-round mechanism is uniquely defined by a combinatorial bandit algorithm which decides the subset selection K_t in each round.

Definition 3. *Bidder i's utility function u_i with respect to bid b_i and other bidders' bids $\mathbf{b}_{-i} = (b_1, \cdots, b_{i-1}, b_{i+1}, \cdots, b_n)$ is*

$$u_i(\mathbf{C}, b_i, \mathbf{b}_{-i}) = v_i \cdot x_i(\mathbf{C}, b_i, \mathbf{b}_{-i}) - p_i(\mathbf{C}, b_i, \mathbf{b}_{-i}).$$

We expect our mechanisms to be truthful, i.e., reporting the real value is the utility-maximizing strategy for any bidder. We define two notions of truthfulness. Ex-post truthfulness requires the mechanism to be truthful on any fixed realization table. Stochastically truthfulness is a weaker notion, which only requires truthfulness in expectation, i.e., the expected utility function is maximized by truthful bidding.

Definition 4. *A mechanism is ex-post truthful if for any $i, v_i, b_i, \mathbf{b}_{-i}, \mathbf{C}$,*

$$u_i(\mathbf{C}, v_i, \mathbf{b}_{-i}) \geq u_i(\mathbf{C}, b_i, \mathbf{b}_{-i}).$$

Definition 5. *A mechanism is stochastically truthful if for any $D, i, v_i, b_i, \mathbf{b}_{-i}$,*

$$\mathbb{E}_{\mathbf{C}}[u_i(\mathbf{C}, v_i, \mathbf{b}_{-i})] \geq \mathbb{E}_{\mathbf{C}}[u_i(\mathbf{C}, b_i, \mathbf{b}_{-i})].$$

The mechanisms should also satisfy IR (Individual Rationality), which ensures that advertisers have non-negative utility when participating in the auction.

Definition 6. *A mechanism is ex-post IR if for any $i, v_i, b_i, \mathbf{b}_{-i}, \mathbf{C}$,*

$$u_i(\mathbf{C}, v_i, \mathbf{b}_{-i}) \geq 0.$$

3 IMPOSSIBILITY RESULT

In this section, we recognize the impossibility for any bandit mechanism to simultaneously achieve sublinear regret and stochastically truthfulness, as shown in Theorem 1. This result reveals that in our two-stage auction setting, it is challenging to design a mechanism that enjoys both good performance and game-theoretic properties.

Theorem 1. *There exists an instance set such that any stochastically truthful algorithm π must incur $\Omega(T)$ regret.*

The proof of Theorem 1 relies on Myerson's Lemma.

Lemma 1 (Myerson's Lemma [23]). *A mechanism is stochastically truthful if and only if any bidder's allocation $\mathbb{E}_{\mathbf{C}}[X_i(\mathbf{C}, b_i, \mathbf{b}_{-i})]$ is monotone(i.e. non-decreasing) with respect to her bid b_i, and the payment rule is given by an explicit formula.*

Now we construct an offline problem instance such that the social welfare-maximizing allocation is not monotone with respect to one's bid.

Proposition 1 (Optimal offline allocation is not monotone). *In the offline (full-information) setting, there exists $D = (D_1, \cdots, D_n)$, and an advertiser i, such that her expected optimal allocation $\mathbb{E}_{\mathbf{r}\sim D}[x_i(\mathbf{r}, \mathbf{b}, K^*)]$, where $K^* = \mathrm{argmax}_K R_D(K, \mathbf{b})$ is the optimal super arm, is not monotone.*

Proof. We present a counterexample with $n = 3$ and $k = 2$. The CTR distributions $D = (D_1, D_2, D_3)$ are

$$c_1 = \begin{cases} 0.8 & \text{with prob. } 0.5 \\ 0.001 & \text{with prob. } 0.5 \end{cases}, c_2 = \begin{cases} 0 & \text{with prob. } 0.7 \\ 1 & \text{with prob. } 0.3 \end{cases}, c_3 = 0.32.$$

Consider bids $\mathbf{b} = (1, 1, 1)$ and $\mathbf{b}' = (1.5, 1, 1)$, where advertiser 1 raises her bid.

On D and $\mathbf{b}$, on can easily compute that the optimal subset is $K^* = \{1,2\}$, and advertiser 1's expected allocation $\mathbb{E}[x_1] = \Pr[c_1 b_1 > c_2 b_2] = 0.7$; while on D and $\mathbf{b}'$, the optimal subset shifts to $K^{*'} = \{1,3\}$ and advertiser 1's expected allocation decreases to $\mathbb{E}[x_1'] = \Pr[c_1 b_1' > c_3 b_3'] = 0.5$. □

In the proof of Proposition 1, the increase of bidder 1's bid leads to a change of the optimal subset, which introduces a stronger competitor (bidder 3) to bidder 1 during the auction within the subset and finally causes the expected allocation of bidder 1 to decrease. This example reveals a fundamental difference between two-stage and one-stage auctions: in two-stage auctions, one bidder may change her competitors in the second stage by changing her bid.

Goel et al. [13] also proved an impossibility result by a construction similar to our Proposition 1. They proved that ex-post truthfulness is not achievable in two-stage auctions, while we present a stronger result, i.e. even stochastically truthfulness is impossible.

To finally prove Theorem 1, we still need the following simple lemma from the bandit literature.

Lemma 2. *Let π be a combinatorial bandit algorithm, and $\mathcal{I} = (D, \mathbf{b})$ be an instance. Assume $\mathcal{I}$ has a unique optimal super arm $K^* = \operatorname{argmax}_K R_D(K, \mathbf{b})$. Let $\tau_K(T) = \sum_{t=1}^{T} \mathbb{I}\{K_t = K\}$ denote the times of π playing K from round 1 to T. If π achieves sublinear regret on $\mathcal{I}$, then $\lim_{T\to\infty} \mathbb{E}[\tau_{K^*}(T)]/T = 1$.*

Now we are ready to prove Theorem 1.

Proof of Theorem 1. We consider two instances $\mathcal{I} = \{D, \mathbf{b}\}$ and $\mathcal{I}' = \{D, \mathbf{b}'\}$, where $D, \mathbf{b}, \mathbf{b}'$ are the constructions in the proof of Proposition 1.

Since algorithm π is stochastically truthful, by Lemma 1, bidder 1's expected allocation must be monotone with respect to b_1. We use shorthand $\mathbb{E}[X_1]$ for $\mathbb{E}[X_1(\mathbf{C}, b_1, \mathbf{b}_{-1})]$, and $\mathbb{E}[X_1']$ for $\mathbb{E}[X_1(\mathbf{C}, b_1', \mathbf{b}'_{-1})]$. Let $\tau_K(T) = \sum_{t=1}^{T} \mathbb{I}\{K_t = K\}$ when π is running on $\mathcal{I}$, and $\tau_K'(T)$ be its counterpart on $\mathcal{I}'$.

We decompose the total allocation $\mathbb{E}[X_1]$ to the times when different super arms containing arm 1 are pulled.

$$\begin{aligned}&\mathbb{E}[X_1]\\ =&\mathbb{E}\left[\tau_{\{1,2\}}(T)\right]\Pr[c_1 b_1 > c_2 b_2] + \mathbb{E}\left[\tau_{\{1,3\}}(T)\right]\Pr[c_1 b_1 > c_3 b_3]\\ =&0.7\mathbb{E}\left[\tau_{\{1,2\}}(T)\right] + 0.5\mathbb{E}\left[\tau_{\{1,3\}}(T)\right].\end{aligned} \tag{1}$$

Similarly,

$$\mathbb{E}[X_1'] = 0.85\mathbb{E}\left[\tau'_{\{1,2\}}(T)\right] + 0.5\mathbb{E}\left[\tau'_{\{1,3\}}(T)\right]. \tag{2}$$

If π achieves sublinear regret on both $\mathcal{I}$ and $\mathcal{I}'$, by applying Lemma 2 to $\mathcal{I}$ and $\mathcal{I}'$, we know that

$$\lim_{T\to\infty} \mathbb{E}\left[\tau_{\{1,2\}}(T)\right]/T = 1 \text{ and } \lim_{T\to\infty} \mathbb{E}\left[\tau'_{\{1,3\}}(T)\right]/T = 1. \tag{3}$$

Moreover,

$$\lim_{T\to\infty} \mathbb{E}\left[\tau_{\{1,3\}}(T)\right]/T = 0 \text{ and } \lim_{T\to\infty} \mathbb{E}\left[\tau'_{\{1,2\}}(T)\right]/T = 0. \tag{4}$$

Combining (3), (4) and (1), (2), we have

$$\lim_{T\to\infty} \mathbb{E}[X_1]/T = 0.7 \text{ and } \lim_{T\to\infty} \mathbb{E}[X_1']/T = 0.5,$$

which contradicts with $\mathbb{E}[X_1] \le \mathbb{E}[X_1']$ and finishes the proof. □

4 TRUTHFUL BANDIT MECHANISMS

In this section, we first introduce truthful approximation oracles, which allows us to design truthful mechanisms at the cost of sacrificing the performance. Then we present two mechanisms utilizing truthful approximation oracles, both of which are ex-post truthful and ex-post IR. The first mechanism achieves $O(T^{2/3})$ regret. The second mechanism requires additional assumptions on the oracle and the bidders' values, and achieves an improved regret of $O(\log T/\Delta_\phi^2)$, where Δ_ϕ is a distribution-dependent gap.

4.1 Truthful Approximation Oracles

Theorem 1 states the impossibly to design a truthful bandit mechanism that approaches the offline optimal allocation. To overcome the difficulty, we introduce approximation oracles that solves the offline optimization problem in a truthful manner.

Definition 7 (Truthful Approximation Oracle). *An oracle takes distributions D and bid vector $\mathbf{b}$ as input, and outputs a subset $K \leftarrow Oracle(D, \mathbf{b})$, $|K| \le k$. For $\alpha \in (0,1)$, an oracle is an $\alpha-$approximation if for any D and any $\mathbf{b}$*

$$R_D(Oracle(D, \mathbf{b})) \ge \alpha \mathrm{OPT}_D.$$

An oracle is ex-post truthful if for any D, $\mathbf{v} \in [0, V]^n, \mathbf{r} \in [0,1]^n$, the following mechanism is truthful:

- *Solicit bid vector $\mathbf{b}$.*
- *Query the oracle for $K \leftarrow Oracle(D, \mathbf{b})$.*
- *Run second-price auction within K. Output $\mathbf{x}(\mathbf{b}, \mathbf{c}, K)$ and $\mathbf{p}(\mathbf{b}, \mathbf{c}, K)$.*

Representation of Distributions. One may wonder how to represent distributions for the oracle's input. In fact, our proposed algorithms only call the oracles on empirical distributions, which are discrete distributions with finite support. We represent such distributions $D = \{D_1, \cdots, D_n\}$ by their CDFs $\mathbf{F} = \{F_1, \cdots, F_n\}$. Each F_i is a piecewise constant function, which could be further represented by a vector of supported points and the values of CDF on those points. Throughout the paper, we may use D and $\mathbf{F}$ interchangeably.

Note that we actually required truthful approximation oracles to be ex-post truthful, rather than the weaker notion of stochastically truthful. The following lemma provides a concrete example of such an ex-post truthful approximation oracle.

Lemma 3 (Theorem 3 in Goel et al. [13]). *For each distribution D_i of c_i, let $\phi_i(\theta)$ be the expectation above the quantile function $q_i(\theta)$:*

$$q_i(\theta) = \sup\{x | \Pr[c_i \ge x] \ge \theta\},$$

$$\phi_i(\theta) = \mathbb{E}_{c_i \sim D_i}[c_i \mathbb{I}[c_i \ge q_i(\theta)]],$$

then the following oracle is a truthful $\frac{e-1}{2e}$-approximation oracle: Sort all bidders by $b_i \phi_i(1/k)$, and choose the top k bidders (ties are broken consistently).

We leverage truthful approximation oracles in our design of truthful online learning algorithms (mechanisms). In such cases, it is not fair to compare our algorithm with the optimal algorithm which always chooses the optimal super arm. Instead, we use the $\alpha-$approximation regret to evaluate an algorithm.

Definition 8. *The $\alpha-$approximation regret of an algorithm is defined as*

$$Reg^{\alpha}(T) = \alpha T \cdot \mathrm{OPT}_D - \mathbb{E}\left[\sum_{t=1}^{T} R_D(K_t)\right].$$

4.2 $O(T^{2/3})$ Mechanism

We present a mechanism that achieves $O(T^{2/3})$ gap-independent regret and ex-post truthfulness. The mechanism consists of a fixed-length exploration phase followed by an exploitation phase. In the exploration phase, we collect m samples for each arm in a round-robin manner, and calculate their empirical distributions in terms of CDFs. With m i.i.d. samples $X_1, \cdots, X_m$ from distribution D, the empirical CDF is defined as $\hat{F}(x) = \frac{1}{m}\sum_{j=1}^{m}\mathbb{I}(X_j \le x)$. Based on these empirical CDFs, a truthful approximation oracle decides the super arm that is repeatedly played in the exploitation phase.

Algorithm 1 An ETC mechanism

```
m ← (π/2)^{1/3} (1+α)^{2/3} n^{2/3} T^{2/3}
for round t ≤ mk do                                  ▷ Exploration Phase
    K_t ← {1 + kt mod n, 2 + kt mod n, ··· , k + kt mod n},
    run second-price auction within K_t
    Update F̂_i for each i ∈ K_t
end for
Query the oracle, get K ← Oracle(F̂, b)
for each remaining round t do                        ▷ Exploitation Phase
    K_t ← K, run second-price auction within K_t
end for
```

Theorem 2. *Algorithm 1 achieves the following $\alpha-$approximation regret upper bound:*

$$Reg^{\alpha}(T) \le 3(1+\alpha)^{\frac{2}{3}} kV n^{\frac{2}{3}} T^{\frac{2}{3}}.$$

The proof of Theorem 2 relies on several useful lemmas.

Lemma 4 (Dvoretzky-Kiefer-Wolfowitz inequality[9, 22]). *Consider a distribution D, and let $F(x)$ be its CDF. With m i.i.d. samples $X_1, \cdots, X_m$ from D, the empirical CDF is $\hat{F}(x) = \frac{1}{m}\sum_{j=1}^{m}\mathbb{I}(X_j \le x)$, then for any $\epsilon > 0$, we have*

$$\Pr[\sup_{x\in R} |F(x) - \hat{F}(x)| \ge \epsilon] \le 2e^{-2m\epsilon^2}$$

Lemma 5. *For random variable X with non-negative support,*

$$\mathbb{E}[X] = \int_0^{\infty} \Pr[X > \epsilon]\mathrm{d}\epsilon.$$

Lemma 6 (Lemma 3 in Chen et al. [5]). *If for any $i \in [n], x \in [0,1]$, $\sup_x |F_i(x) - F_i'(x)| \le \Lambda$, then for any super arm K, we have $|R_F(K) - R_{F'}(K)| \le 2Vk\Lambda$.*

Lemma 4, i.e. the DKW inequality, is on concentration of empirical distributions. Lemma 5 is a simple fact in probability theory, bridging expectations and CDFs. Lemma 6 characterizes the concentration of super arms' rewards based on the concentration of base arms' empirical distributions. Now we are ready to prove Theorem 2.

Proof. Let K be the super arm played in the exploitation phase. Let $\hat{D} = (\hat{D}_1, \cdots, \hat{D}_n)$ be the empirical distributions at round mk, with empirical CDFs $\hat{\mathbf{F}} = (\hat{F}_1, \cdots, \hat{F}_n)$, and let D be the real distributions. Since we call an $\alpha-$approximation oracle, we have $R_{\hat{D}}(K) \ge \alpha\mathrm{OPT}_{\hat{D}}$. Let $K^* = \mathrm{argmax}_K R_D(K)$ be the real optimal super arm. Then

$$R_{\hat{D}}(K) \ge \alpha\mathrm{OPT}_{\hat{D}} \ge \alpha R_{\hat{D}}(K^*),$$

where the second inequality follows from the optimality of $\mathrm{OPT}_{\hat{D}}$, i.e., $\mathrm{OPT}_{\hat{D}} \ge R_{\hat{D}}(K')$ for any K'. We then bound the probability $\Pr[R_D(K) < \alpha\mathrm{OPT}_D - \epsilon]$ for any $\epsilon \in \mathbb{R}^+$.

$$\begin{aligned}
&\Pr[R_D(K) < \alpha\mathrm{OPT}_D - \epsilon] \\
\le&\Pr[R_D(K) < \alpha\mathrm{OPT}_D - \epsilon + (R_{\hat{D}}(K) - \alpha R_{\hat{D}}(K^*))] \\
=&\Pr[(R_{\hat{D}}(K) - R_D(K)) - \alpha(R_{\hat{D}}(K^*) - R_D(K^*)) > \epsilon] \qquad (5)\\
\le&\Pr[|R_{\hat{D}}(K) - R_D(K)| + \alpha|R_{\hat{D}}(K^*) - R_D(K^*)| > \epsilon] \\
:=&\Pr[\mathcal{E}],
\end{aligned}$$

where the last inequality follows from $|a-b| \le |a| + |b|$ for any $a, b \in \mathbb{R}$, and in the last line we define event $\mathcal{E} = \{|R_{\hat{D}}(K) - R_D(K)| + \alpha|R_{\hat{D}}(K^*) - R_D(K^*)| > \epsilon\}$.

Also, define good event

$$\mathcal{G} = \left\{\forall i \in [n], \sup_{x\in[0,1]} |F_i(x) - \hat{F}_i(x)| \le \frac{\epsilon}{2Vk(1+\alpha)}\right\}.$$

By Lemma 6, if $\mathcal{G}$ happens, then we have $|R_{\hat{D}}(K) - R_D(K)| \le \frac{\epsilon}{1+\alpha}$ and $|R_{\hat{D}}(K^*) - R_D(K^*)| \le \frac{\epsilon}{1+\alpha}$, which prevents $\mathcal{E}$ to happen. Therefore, $\mathcal{E}$ implies $\neg\mathcal{G}$, which gives us

$$\begin{aligned}
&\Pr[\mathcal{E}] \\
\le&\Pr[\neg\mathcal{G}] \\
=&\Pr\left[\exists i \in [n], \sup_{x\in[0,1]} |F_i(x) - \hat{F}_i(x)| > \frac{\epsilon}{2Vk(1+\alpha)}\right] \\
\le&\sum_{i=1}^{n}\Pr\left[\sup_{x\in[0,1]} |F_i(x) - \hat{F}_i(x)| > \frac{\epsilon}{2Vk(1+\alpha)}\right] \qquad (6)\\
\le&2n\exp\left(-\frac{m\epsilon^2}{2V^2k^2(1+\alpha)^2}\right),
\end{aligned}$$

where the second inequality is by taking a union bound, and the third inequality follows from DKW inequality.

Combining (5) and (6) gives us

$$\Pr\left[R_D(K) < \alpha\mathrm{OPT}_D - \epsilon\right] \le 2n\exp\left(-\frac{m\epsilon^2}{2V^2k^2(1+\alpha)^2}\right).$$

Split the regret by exploration phase and exploitation phase,

$$Reg^{\alpha}(T) = mkV + (T - mk)\mathbb{E}[\alpha\mathrm{OPT}_D - R_D(K)]. \qquad (7)$$

Calculate the expectation term with Lemma 5,

$$\begin{aligned}
&\mathbb{E}[\alpha\mathrm{OPT}_D - R_D(K)] \\
=&\int_0^{\infty}\Pr[\alpha\mathrm{OPT}_D - R_D(K) > \epsilon]\mathrm{d}\epsilon \\
\le&\int_0^{\infty} 2n\exp\left(-\frac{2m\epsilon^2}{4V^2k^2(1+\alpha)^2}\right)\mathrm{d}\epsilon \qquad (8)\\
=&\frac{\sqrt{2\pi}nVk(1+\alpha)}{\sqrt{m}}.
\end{aligned}$$

Plug (8) into (7) , and let $m = \min(\lceil (\frac{\pi}{2})^{\frac{1}{3}}(1+\alpha)^{\frac{2}{3}} n^{\frac{2}{3}} T^{\frac{2}{3}} \rceil, \lceil T/k \rceil)$, we have

$$\begin{aligned} Reg^{\alpha}(T) =& mkV + (T - mk)\mathbb{E}[\alpha \mathrm{OPT}_D - R_D(K)] \\ \leq& mkV + T\frac{\sqrt{2\pi} nVk(1+\alpha)}{\sqrt{m}} \\ =& (2^{-\frac{2}{3}} + 2^{-\frac{1}{3}})(2\pi)^{\frac{2}{3}}(1+\alpha)^{\frac{2}{3}} kVn^{\frac{2}{3}} T^{\frac{2}{3}} \\ \leq& 3(1+\alpha)^{\frac{2}{3}} kVn^{\frac{2}{3}} T^{\frac{2}{3}}. \end{aligned} \tag{9}$$

□

Beyond sublinear regret, Algorithm 1 also enjoys the following truthful and IR properties.

Proposition 2. *Algorithm 1 is ex-post truthful and ex-post IR.*

The truthfulness of Algorithm 1 mainly follows from the property of truthful oracles combined with the truthfulness of second-price auctions. The IR property follows from that of second-price auctions. The complete proofs are deferred to Appendix A.

4.3 $O(\log T/\Delta_{\phi}^2)$ Mechanism

Although Algorithm 1 guarantees ex-post truthfulness and ex-post IR, its $O(T^{2/3})$ regret is not satisfactory in some scenarios. To design an algorithm with a better regret bound, we make an additional assumption on the oracle. Beyond truthfulness and $\alpha-$approximation, we assume that the oracle determines K by ranking the bidders according to some score and choosing the top-k, and the scores are accessible. The scores provide additional information about the quality of the bidders, and can be leveraged by the bandit algorithm to quickly determine the correct subset to exploit.

Definition 9 (Truthful Approximation Scoring Oracle). *A scoring oracle Score$(\cdot)$ assigns each bidder a score according to its distribution, i.e., $\phi_i \leftarrow$ Score(D_i), such that ranking the bidders by $b_i\phi_i$, and choosing the top-k as K will ensure*

$$R_D(K) \geq \alpha \mathrm{OPT}_D.$$

It is easy to check that this scoring and ranking procedure is always ex-post truthful for any score $\phi = (\phi_1, \cdots, \phi_n)$.

We further make an assumption on the smoothness of Score$(\cdot)$: There exists a Lipschitz constant $L > 0$, such that if $\sup_x |F(x) - F'(x)| \leq \Lambda$, then

$$|\mathrm{Score}(F) - \mathrm{Score}(F')| \leq L\Lambda.$$

The following lemma tells us such scoring oracle exists. In fact, the oracle presented in Lemma 3 is exactly a truthful $\frac{e-1}{2e}-$ approximation scoring oracle. The proof of its Lipschitz constant $L = 1$ is deferred to Appendix A.

Lemma 7. *There exists a truthful $\frac{e-1}{2e}$-approximation scoring oracle with Lipschitz constant $L = 1$.*

We denote the gap of an instance as $\Delta_{\phi} = \min_{i,j\in[n],i\neq j} |v_i\phi_i - v_j\phi_j|$. For the truthful property of the mechanism, we assume that the value v_i of each arm i is within an interval around a known parameter $\beta_i \in [0, V]$:

$$v_i \in \left[\beta_i - \Delta_{\phi}/2\phi_i, \beta_i + \Delta_{\phi}/2\phi_i\right]. \tag{10}$$

This assumption is often satisfied in industrial scenarios where advertisers' valuation of a certain impression is static, and the ad platform can estimate one's value from one's bidding history.

Based on an $\alpha-$approximation scoring oracle, we design an ETC mechanism with adaptive commitment time, i.e., the length of exploration phase depends on the collected data. The fixed prior information of value β_i, instead of submitted bid b_i, is used in deciding of commitment. This prevents the bidders from influencing the commitment time by strategic bidding.

Algorithm 2 An ETC mechanism with adaptive commitment time

Throughout the exploration phase, for each arm $i \in [n]$ we maintain: (i) counter T_i which stores the times that i has been selected so far (ii) CDF $\hat{F}_i$ of the empirical distribution of the observed outcomes of arm i so far
repeat ▷ Exploration Phase
  $K_t \leftarrow \{1 + kt \mod n, 2 + kt \mod n, \cdots, k + kt \mod n\}$, run second-price auction within K_t
  for each $i \in K_t$ **do**
    Update T_i and $\hat{F}_i$
    Query the scoring oracle, get $\hat{\phi}_i \leftarrow \mathrm{Score}(\hat{F}_i)$, compute $\overline{\phi_i} \leftarrow \hat{\phi}_i + L\sqrt{\frac{\log T}{T_i}}$ and $\underline{\phi_i} \leftarrow \hat{\phi}_i - L\sqrt{\frac{\log T}{T_i}}$
  end for
  Rank all the arms by $\beta_i\hat{\phi}_i$, let $High$ be the set of top-k arms, let Low be the other $n - k$ arms
  $\phi_h \leftarrow \min_{i\in High} \beta_i\underline{\phi_i}$, $\phi_l \leftarrow \max_{i\in Low} \beta_i\overline{\phi_i}$
until $\phi_h \geq \phi_l$
for each remaining round t **do** ▷ Exploitation Phase
  Pull $K_t \leftarrow High$, run second-price auction within K_t, using bids $\mathbf{b}$
end for

Theorem 3. *Algorithm 2 achieves the following $\alpha-$approximation regret upper bound:*

$$Reg^{\alpha}(T) \leq 2nV + \frac{4nL^2V^3}{k\Delta_{\phi}^2}\log T.$$

Proof. Let $\hat{F}_{i,u}(x)$ be the empirical distribution of arm i when u samples from i are observed. Define event

$$\mathcal{E} = \left\{\exists i \in [n], \exists u \in [T], \sup_{x\in[0,1]} \left|\hat{F}_{i,u}(x) - F_i(x)\right| \geq \sqrt{\frac{\log T}{u}}\right\}.$$

From the DKW inequality, for $\forall i \in [n], \forall u \in [T]$,

$$\Pr\left[\sup_{x\in[0,1]} |\hat{F}_{i,u}(x) - F_i(x)| \geq \sqrt{\frac{\log T}{u}}\right] \leq 2e^{-2u\frac{\log T}{u}} = \frac{2}{T^2}.$$

Taking an union bound gives $\Pr[\mathcal{E}] \leq \frac{2n}{T}$.

Let $\hat{\phi}_{i,t}, \overline{\phi}_{i,t}, \underline{\phi}_{i,t}$ be the value of $\hat{\phi}_i, \overline{\phi}_i, \underline{\phi}_i$ at time t. On event $\neg\mathcal{E}$, by definition, for any arm i, any time t in the exploration phase, $|\phi_i - \hat{\phi}_{i,t}| \leq L\sqrt{\frac{\log T}{T_i}}$, therefore $\underline{\phi}_{i,t} \leq \phi_i \leq \overline{\phi}_{i,t}$. When condition $\phi_h \geq \phi_l$ is satisfied, for $\forall i \in High, \forall j \in Low$, $\beta_i\underline{\phi}_{i,t} \geq \beta_j\overline{\phi}_{j,t}$, thus $\beta_i\phi_i \geq \beta_j\phi_j$. This implies that $High$ is the top-k subset according

to $\{\beta_i\phi_i\}$. By Equation (10), $High$ is also the top-k subset according to $\{b_i\phi_i\}$.

Let m be the last round in which $\phi_h < \phi_l$. Let $i = \text{argmin}_{i\in High}\, \beta_i\underline{\phi}_i$, $j = \text{argmax}_{j\in Low}\, \beta_j\overline{\phi}_j$, we know $\beta_i\underline{\phi}_i < \beta_j\overline{\phi}_j$. By the exploration rule, till round m, we have observed $\frac{mk}{n}$ samples for each arm. Since $\beta_i\underline{\phi}_i < \beta_j\overline{\phi}_j$,

$$\beta_i\left(\phi_i - L\sqrt{\frac{n\log T}{mk}}\right) < \beta_j\left(\phi_j + L\sqrt{\frac{n\log T}{mk}}\right).$$

Rearranging the terms, we have

$$\Delta_\phi \le \beta_i\phi_i - \beta_j\phi_j \le (\beta_i+\beta_j)L\sqrt{\frac{n\log T}{mk}} \le 2LV\sqrt{\frac{n\log T}{mk}}.$$

Solving m from the first and last term gives $m \le \frac{4nL^2V^2\log T}{k\Delta_\phi^2}$.

On event $\mathcal{E}$, each round incurs regret of at most V. On event $\neg\mathcal{E}$, we know that $High$ is the top-k subset according to $\{b_i\phi_i\}$. By definition of the $\alpha-$approximation scoring oracle, playing $High$ incurs non-positive regret. Therefore, the regret of the algorithm is

$$\begin{aligned} Reg^\alpha(T) &\le \Pr[\mathcal{E}]TV + \Pr[\neg\mathcal{E}]\mathbb{E}[m|\neg\mathcal{E}]V \\ &\le \frac{2n}{T}TV + \frac{4nL^2V^2\log T}{k\Delta_\phi^2}V \\ &= 2nV + \frac{4nL^2V^3}{k\Delta_\phi^2}\log T. \end{aligned}$$

□

Algorithm 2 also achieves ex-post truthfulness and ex-post IR. The proofs are deferred to Appendix A.

Proposition 3. *Algorithm 2 is ex-post truthful and ex-post IR.*

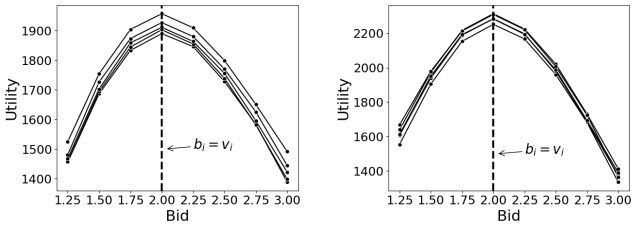
By the design of adaptive commitment time, Algorithm 2 achieves a better regret than Algorithm 1 on most instances, except for the cases when Δ_ϕ is extremely small. A potential way to further improve the regret bound is to adopt UCB-based [5] or successive-elimination style [27] algorithms. However, these algorithms are much more data sensitive than ETC algorithms, thus may be prone to strategic bids.

5 EXPERIMENTS

In this section, we evaluate our two mechanisms through experiments on both synthetic data and real-world data.

For the experiments, instead of $\alpha-$approximation regret (Definition 8), we compare the cumulative rewards achieved by our mechanisms against $T\cdot R_D(K_{\text{Oracle}})$, where K_{Oracle} is the subset returned by an $\alpha-$approximation oracle. By definition of $\alpha-$approximation oracles (Definition 7), $R_D(K_{\text{Oracle}}) \ge \alpha\text{OPT}_D$, so $R_D(K_{\text{Oracle}})$ is a more challenging reference value, and all of our theoretical results still holds under this notion of regret. This choice is based on two reasons: (i) computing OPT_D is often computationally infeasible, (ii) the value of αOPT_D can be much lower than $R_D(K_{\text{Oracle}})$ practically, and comparing the mechanisms' cumulative reward against αOPT_D often leads to negative regret.

Beyond regret, we also test the truthfulness of our mechanisms, by computing a bidder's utility with different bids.



(a) Ex-post truthfulness of Algorithm 1. **(b) Ex-post truthfulness of Algorithm 2.**

Figure 1: Ex-post truthfulness of our two algorithms evaluated on synthetic data. Each line represents a bidder's utilities with respect to different submitted bids on one random seed.

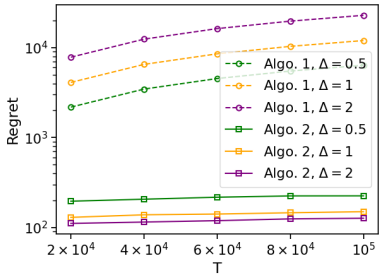
5.1 Experiments with synthetic data

5.1.1 Experiment Setup. We construct an environment with $n = 7$ bidders, from which $k = 3$ bidders are selected in each round. All bidders have the same uniform CTR distribution, i.e., $c_i \sim U([0,1])$. The bidders' values are $[1+\Delta, 1+\Delta, 1+\Delta, 1, 1, 1, 1]$, where Δ is a gap parameter that we control through the experiments. By the truthful property of our mechanisms, the input bids are equal to the values. We run experiments for different time horizons $T \in \{2\times10^4, 4\times10^4, 6\times10^4, 8\times10^4, 10^5\}$. For each time horizon, we run experiments for $\Delta \in \{0.5, 1, 1.5, 2\}$. The result of each experiment is averaged over 80 independent runs.

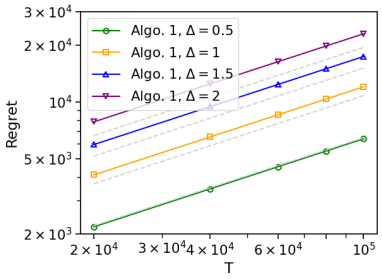
We leverage the oracle presented in 3 for both Algorithm 1 and Algorithm 2. The difference is that for Algorithm 1, the oracle only returns a subset it chooses, while for Algorithm 2, it returns all the scores $\phi_i(1/k)$ for $i \in [n]$.

To test the ex-post truthfulness of our mechanisms, we fix Δ to be 1, so bidders' the values are $[2, 2, 2, 1, 1, 1, 1]$. Since ex-post truthfulness requires the mechanism to be truthful on any random seed, we pick 100 random seeds for evaluation. For each fixed random seed, we adjust the bid of bidder 1 to $b_1 \in \{1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00\}$, while keeping other bidders' bids unchanged. We report bidder 1's utility when different bids are reported. If the mechanism is ex-post truthful, then for any random seed, the utility-maximizing bid should be equal to the value. During the test of truthfulness, we fix $T = 10000$.

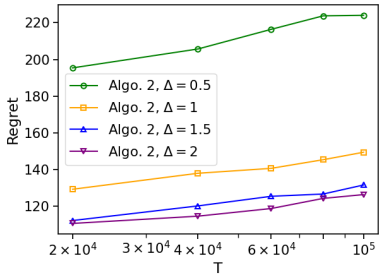
5.1.2 Results and Discussions. Figure 2a presents a comparison of regret between Algorithm 1 and Algorithm 2. For any time horizon T and gap Δ, the regret of Algorithm 2 is significantly lower than that of Algorithm 1. The low regret is due to the design of adaptive commitment time in Algorithm 2. Besides, when the gap Δ decreases, Algorithm 1 achieves lower regret, while Algorithm 2 suffers from higher regret. This phenomenon is demonstrated more clearly in Figure 2b and Figure 2c. Figure 2b depicts the regret of Algorithm 1 with different T and Δ in a log-log plot. The grey dashed lines represent $Reg = aT^{\frac{2}{3}}$ with different values of a. We observe that the regret curves are almost parallel with the grey lines, which indicates $\Theta(T^{2/3})$ regret, matching Theorem 2. Moreover, Algorithm 1 shows higher regret when Δ gets high. This is because the regret accumulated in the exploration phase grows as the gap between optimal and suboptimal super arms expands.



(a) Comparison of the regret of two mechanisms. Regret is displayed on a log scale.



(b) The regret of Algorithm 1. To demonstrate the order of $T^{2/3}$, both Regret and T are displayed on a log scale.



(c) The regret of Algorithm 2. To demonstrate the order of $\log T$, T is displayed on a log scale.

Figure 2: Regret of our two algorithms evaluated on synthetic data.

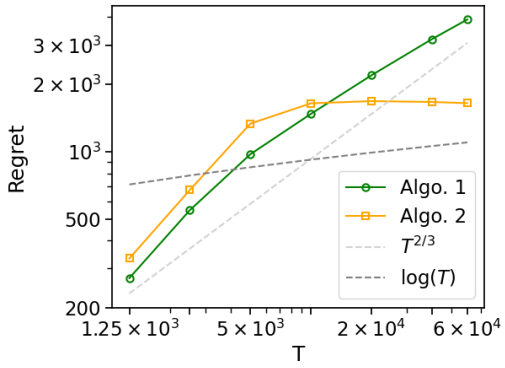


Figure 3: Regret of our two algorithms evaluated on MovieLens dataset. Both Regret and T are displayed on a log scale.

Figure 2c shows the regret of Algorithm 2. We observe that the regret almost grows linearly with respect to $\log T$. Besides, the regret grows quadratically with decreasing Δ. Note that in the case of our experiment, Δ is proportional to Δ_ϕ, as the distributions are identical and fixed. The dependence of regret on T and Δ_ϕ nicely matches our theoretical result (Theorem 3).

In the test of truthfulness, on all 100 random seeds, bidding the true value achieves highest utility among 8 different bids. Figure 1 shows the utility curve of two mechanisms on five random seeds $\{1, 2, 3, 4, 5\}$. The curves on all 100 seeds actually look similar, with $b = 2$ being their common maximum point, and the fives seeds are arbitrarily picked only for demonstration.

5.2 Experiments with real-world data

We evaluate the Algorithm 1 and Algorithm 2 on the MovieLens 1M [14] dataset.

5.2.1 Experiment Setup. We treat each movie as an arm and convert the ratings into a CTR-like metric. The top 7 arms, determined by the highest number of ratings in the original dataset, are selected as the base arms. From these base arms, a super arm consisting of $k = 3$ arms is chosen in each round. To construct the CTR distribution, we manually set a rating threshold $r_t = 3.5$, such that ratings $r > 3.5$ are converted to 1, and otherwise to 0. Subsequently, we calculate the mean and variance of the converted ratings for each arm, denoted as the mean and variance of the CTR distribution, respectively. These CTR distributions are modeled as truncated Gaussian distributions, with support restricted to $[0, 1]$. The bid of each arm is sampled from a uniform distribution $U([0, 5])$ and remains constant after the experiment begins. We conduct experiments for different time horizons $T \in \{1250, 2500, 5000, 1 \times 10^4, 2 \times 10^4, 4 \times 10^4, 6 \times 10^4\}$. The setting of the oracle is the same as Section 5.1.1.

5.2.2 Results and Discussions. Figure 3 presents a comparison of the regret between Algorithm 1 and Algorithm 2. When the horizon T is small, both algorithms exhibit linear regret, as $mk > T$ for a small T, leading to a predominantly exploratory phase within the limited horizon. As T increases, both algorithms demonstrate improvements by incorporating an exploitation phase. In comparison with the dashed grey line, it is evident that Algorithm 1 achieves $O(T^{2/3})$ regret, while Algorithm 2 attains $O(\log T)$ regret. Notably, with a large horizon T, Algorithm 2 achieves lower regret compared to Algorithm 1.

6 RELATED WORKS

Two-stage Advertising Systems. Previous studies on two-stage advertising systems have primarily focused on two aspects: allocation efficiency and incentives. Both Wang et al. [25] and Zhao et al. [32] addressed the learning objectives of machine learning models in the first stage, in order to align with the second stage and enhancing the overall ad allocation performance. While Wang et al. [25] also discussed incentives, they considered a non-standard value-maximizer utility model. On incentives in two-stage auctions, Goel et al. [13] provided an insightful characterization of first-stage mechanisms that ensures overall truthfulness when composed with any truthful second-stage auction mechanism. However, their work was limited to single-round mechanisms, whereas we considered incentives in multi-round bandit mechanisms.

Two-stage Recommendation Systems. Research on two-stage structures in recommendation is rather abundant than advertising. A general approach is cooperative training of both stages [12, 15, 17, 18] to improve the overall recommendation performance, and particular attention has been paid to off-policy correction [20] and fairness issues [24]. Closer to our setting is synchronized two-stage exploration, studied under both linear [16] and neural [31] bandit settings. Although these studies share some similarities with our setting, there are significant differences due to the presence of incentives in ad auctions.

Truthful Bandit Mechanisms. A line of research has focused on designing truthful mechanisms for multi-round ad auctions, where a multi-armed bandit algorithm acts as the allocation rule. Babaioff et al. [4] provided a $\Omega(T^{2/3})$ regret lower bound for any deterministic truthful multi-armed bandit mechanisms, and provided a ETC algorithm that matches this lower bound. Devanur and Kakade [8] obtained a similar $\Omega(T^{2/3})$ lower bound under the revenue-maximizing setting. Babaioff et al. [3] further extended to randomized mechanisms, and provided a black-box reduction from any monotone bandit algorithm to a truthful mechanism, which gives rise to $O(\sqrt{T})$ regret. Recent works have considered extended settings in different directions, such as utility models [11] and contextual information [1, 28, 30]. Our work is based on a novel setting of two-stage ad auctions which is formulated as designing combinatorial bandit mechanisms.

7 CONCLUSION

In this paper, we investigate the problem of designing truthful bandit mechanisms for two-stage online ad auctions. We prove an $\Omega(T)$ lower bound for truthful mechanisms, and introduce truthful $\alpha-$approximation oracles which give rise to sublinear $\alpha-$approximation regret mechanisms.

We leave it as an open problem to potentially design $O(\sqrt{T})$ bandit mechanisms within the approximation regret setting, or to establish an $\Omega(T^{2/3})$ lower bound. Moreover, the impossibility result may also be circumvented by other approaches, e.g., relaxing the notion of truthfulness by considering high-probability truthful mechanisms.

ACKNOWLEDGMENTS

The authors sincerely thank Shuai Li and Xutong Liu for their helpful discussions. This work was supported in part by National Key R&D Program of China (No. 2022ZD0119100), in part by China NSF grant No. 62322206, 62132018, U2268204, 62025204, 62272307, 62372296. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.

REFERENCES

[1] Kumar Abhishek, Shweta Jain, and Sujit Gujar. 2020. Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1732–1734.

[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*. IEEE, 322–331.

[3] Moshe Babaioff, Robert D Kleinberg, and Aleksandrs Slivkins. 2015. Truthful mechanisms with implicit payment computation. *Journal of the ACM (JACM)* 62, 2 (2015), 1–37.

[4] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2014. Characterizing Truthful Multi-armed Bandit Mechanisms. *SIAM J. Comput.* 43, 1 (2014), 194–230.

[5] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. 2016. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems* 29 (2016).

[6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[7] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[8] Nikhil R Devanur and Sham M Kakade. 2009. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*. 99–106.

[9] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* (1956), 642–669.

[10] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*. 1775–1784.

[11] Zhe Feng, Christopher Liaw, and Zixin Zhou. 2023. Improved online learning algorithms for CTR prediction in ad auctions. In *International Conference on Machine Learning*. PMLR, 9921–9937.

[12] Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J. Shane Culpepper. 2019. Joint Optimization of Cascade Ranking Models. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. 15–23.

[13] Gagan Goel, Renato Paes Leme, Jon Schneider, David Thompson, and Hanrui Zhang. 2023. Eligibility Mechanisms: Auctions Meet Information Retrieval. In *Proceedings of the ACM Web Conference 2023*. 3541–3549.

[14] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[15] Jiri Hron, Karl Krauth, Michael Jordan, and Niki Kilbertus. 2021. On Component Interactions in Two-Stage Recommender Systems. In *Advances in Neural Information Processing Systems*, Vol. 34. 2744–2757.

[16] Jiri Hron, Karl Krauth, Michael I. Jordan, and Niki Kilbertus. 2020. Exploration in two-stage recommender systems. arXiv:2009.08956 [cs.IR]

[17] Xu Huang, Defu Lian, Jin Chen, Liu Zheng, Xing Xie, and Enhong Chen. 2023. Cooperative Retriever and Ranker in Deep Recommenders. In *Proceedings of the ACM Web Conference 2023*. 1150–1161.

[18] Wang-Cheng Kang and Julian McAuley. 2019. Candidate Generation with Binary Codes for Large-Scale Top-N Recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. 1523–1532.

[19] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press, Chapter 6, 97–98.

[20] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H. Chi. 2020. Off-policy Learning in Two-stage Recommender Systems. In *Proceedings of The Web Conference 2020*. 463–473.

[21] Xu Ma, Pengjie Wang, Hui Zhao, Shaoguo Liu, Chuhan Zhao, Wei Lin, Kuang-Chih Lee, Jian Xu, and Bo Zheng. 2021. Towards a Better Tradeoff between Effectiveness and Efficiency in Pre-Ranking: A Learnable Feature Selection based Approach. 2036–2040.

[22] Pascal Massart. 1990. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability* (1990), 1269–1283.

[23] Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research* 6, 1 (1981), 58–73.

[24] Lequn Wang and Thorsten Joachims. 2023. Uncertainty Quantification for Fairness in Two-Stage Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 940–948.

[25] Yiqing Wang, Xiangyu Liu, Zhenzhe Zheng, Zhilin Zhang, Miao Xu, Chuan Yu, and Fan Wu. [n. d.]. On Designing a Two-stage Auction for Online Advertising. In *Proceedings of the ACM Web Conference 2022*. 90–99.

[26] Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2020. COLD: Towards the Next Generation of Pre-Ranking System. arXiv:2007.16122 [cs.IR]

[27] Haike Xu and Jian Li. 2021. Simple combinatorial algorithms for combinatorial bandits: Corruptions and approximations. In *Uncertainty in Artificial Intelligence*. PMLR, 1444–1454.

[28] Yinglun Xu, Bhuvesh Kumar, and Jacob Abernethy. 2023. On the robustness of epoch-greedy in multi-agent contextual bandit mechanisms. *arXiv preprint arXiv:2307.07675* (2023).
[29] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 269–277.
[30] Mengxiao Zhang and Haipeng Luo. 2023. Online Learning in Contextual Second-Price Pay-Per-Click Auctions. *arXiv preprint arXiv:2310.05047* (2023).
[31] Mengyan Zhang, Thanh Nguyen-Tang, Fangzhao Wu, Zhenyu He, Xing Xie, and Cheng Soon Ong. 2022. Two-Stage Neural Contextual Bandits for Personalised News Recommendation. arXiv:2206.14648 [cs.IR]
[32] Zhishan Zhao, Jingyue Gao, Yu Zhang, Shuguang Han, Siyuan Lou, Xiang-Rong Sheng, Zhe Wang, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. COPR: Consistency-Oriented Pre-Ranking for Online Advertising. arXiv:2306.03516 [cs.IR]
[33] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
[34] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

A PROOFS

Proof of Lemma 2. Decompose the regret to super arms,

$$\begin{aligned} Reg(T) &= T \cdot \mathrm{OPT}_D - \mathbb{E}\left[\sum_{t=1}^{T} R_D(K_t)\right] \\ &= \sum_{K} (R_D(K^*) - R_D(K))\mathbb{E}\left[\tau_K(T)\right]. \end{aligned}$$

By sublinear regret,

$$0 = \lim_{T\to\infty} \frac{Reg(T)}{T} = \sum_{K} (R_D(K^*) - R_D(K)) \lim_{T\to\infty} \frac{\mathbb{E}[\tau_K(T)]}{T}.$$

Since the optimal super arm is unique, for every $K \neq K^*$, $R_D(K^*) - R_D(K) > 0$, thus we have $\lim_{T\to\infty} \mathbb{E}[\tau_K(T)]/T = 0$. Note that $T = \sum_K \mathbb{E}[\tau_K(T)]$. Therefore $\lim_{T\to\infty} \mathbb{E}[\tau_{K^*}(T)]/T = 1$.

□

Proof of Proposition 2. For truthfulness, fix any realization table $\mathbf{C}$, and consider the cumulative utility $U_i(\mathbf{C}, b_i, \mathbf{b}_{-i})$ of bidder i. We separate U_i to the cumulative utility in exploration phase and exploitation phase,

$$U_i(\mathbf{C}, b_i, \mathbf{b}_{-i}) = \sum_{t=1}^{mk} u_{it}(\mathbf{C}, b_i, \mathbf{b}_{-i}) + \sum_{t=mk+1}^{T} u_{it}(\mathbf{C}, b_i, \mathbf{b}_{-i})$$

where $u_{it}(\mathbf{C}, b_i, \mathbf{b}_{-i})$ is bidder i's utility in round t.

In the exploration phase, bidder i only participates in the subset auction in certain fixed rounds, and its competitors in these rounds is not influenced by b_i. For each of these rounds, by the truthfulness of second-price auctions, bidder i optimizes its utility with truthful bid v_i.

For the exploitation phase, the empirical distributions $\hat{\mathbf{F}}$ in line 6 only depend on $\mathbf{C}$, and are not influenced by b_i. The exploitation phase consists of repeated second-price auctions on a fixed subset K, which is selected by the oracle. From the truthfulness of the oracle, for fixed $\hat{\mathbf{F}}$, the combination of selecting K and running second-price auction within K is truthful, with respect to any realization $\mathbf{c}$. This implies that the combination of selecting K and running any one of the auctions in rounds $mk + 1 \le t \le T$ is truthful. Therefore v_i is the maximizer of any of the objectives $u_{it}(\mathbf{C}, b_i, \mathbf{b}_{-i})$ for $mk + 1 \le t \le T$.

To conclude, the truthful bid v_i maximizes any of $u_{it}(\mathbf{C}, b_i, \mathbf{b}_{-i})$ for $1 \le t \le T$, and thus maximizes $U_i(\mathbf{C}, b_i, \mathbf{b}_{-i})$.

For IR, the total utility U_i of bidder i is defined as the sum of utility in each round,

$$U_i(\mathbf{C}, v_i, \mathbf{b}_{-i}) = \sum_{t=1}^{T} u_{it}(\mathbf{C}, v_i, \mathbf{b}_{-i})$$

Fix any realization table $\mathbf{C}$. For any t, if $i \in K_t$, then i participates a second-price auction in round t. By the IR property of second-price auctions, $u_{it}(\mathbf{C}, v_i, \mathbf{b}_{-i}) \ge 0$. If $i \notin K_t$, both allocation x_{it} and payment p_{it} are zero, thus $u_{it} = 0$. Since each round produces non-negative utility when truthful bidding, the total utility $U_i(\mathbf{C}, v_i, \mathbf{b}_{-i})$ is non-negative, therefore the mechanism is IR. □

Proof of Lemma 7. The oracle in Lemma 3 is a truthful $\frac{e-1}{2e}$-approximation scoring oracle. Now we prove that its Lipschitz constant is $L = 1$.

For distributions D and D', with CDFs F and F',

$$\begin{aligned} \mathrm{Score}(F) &= \phi\left(\frac{1}{k}\right) \\ &= \mathbb{E}_{r\sim D}\left[r \cdot \mathbb{I}\left[r \ge q(\frac{1}{k})\right]\right] \\ &= \int_{q(\frac{1}{k})}^{1} r \mathrm{d}F(r) \\ &= \int_{0}^{q(\frac{1}{k})} \frac{1}{k}\mathrm{d}r + \int_{q(\frac{1}{k})}^{1} (1 - F(r))\mathrm{d}r \end{aligned}$$

Define

$$z(r) := \begin{cases} \frac{1}{k} & 0 \le r < q\left(\frac{1}{k}\right) \\ 1 - F(r) & q\left(\frac{1}{k}\right) \le r \le 1 \end{cases},$$

we have $\mathrm{Score}(F) = \int_0^1 z(r)\mathrm{d}r$.

For the difference of scores,

$$|\mathrm{Score}(F) - \mathrm{Score}(F')| \le \int_0^1 |z(r) - z'(r)|\mathrm{d}r \tag{11}$$

Now we prove that for any $r \in [0, 1]$,

$$|z(r) - z'(r)| \le |F(r) - F'(r)|. \tag{12}$$

Without loss of generality, assume $q(\frac{1}{k}) \le q'(\frac{1}{k})$. Consider three cases:

Case 1. $r < q(\frac{1}{k})$. In this case $z(r) = z'(r) = \frac{1}{k}$, $z(r) - z'(r) = 0$.

Case 2. $q(\frac{1}{k}) \le r < q'(\frac{1}{k})$. By $q(\frac{1}{k}) \le q'(\frac{1}{k})$ we know $F'(r) < 1 - \frac{1}{k}$. Thus $z(r) - z'(r) = 1 - F(r) - \frac{1}{k} \le F'(r) - F(r)$

Case 3. $r \ge q'(\frac{1}{k})$. In this case $z(r) - z'(r) = F'(r) - F(r)$

Concluding the three cases finishes the proof of (12). Plugging (12) into (11) gives us

$$|\mathrm{Score}(F) - \mathrm{Score}(F')| \le \int_0^1 |F(r) - F'(r)|\mathrm{d}r \le \int_0^1 \Lambda \mathrm{d}r \le \Lambda$$

□

Proof of Proposition 3. For truthfulness, fix any realization table $\mathbf{C}$. In the exploration phase, the selected subset do not depend on the bids. Moreover, the output super arm $High$ is also not influenced by the bids. In the exploitation phase, we repeatedly run second-price auctions within $High$. Since a bidder cannot influence the selected subset K_t in any round t, the truthfulness of our mechanism follows from the truthfulness of second-price auction.

For IR, again fix any realization table $\mathbf{C}$. For any round t, and any bidder i, if $i \in K_t$, then i participates a second-price auction in round t. By the IR property of second-price auctions, $u_{it}(\mathbf{C}, v_i, \mathbf{b}_{-i}) \geq 0$. If $i \notin K_t$, both allocation x_{it} and payment p_{it} are zero, thus $u_{it} = 0$. Since each round produces non-negative utility when truthful bidding, the total utility $U_i(\mathbf{C}, v_i, \mathbf{b}_{-i})$ is non-negative. □