Master's Thesis of Data Science

Efficient Exploration for Online Advertising Auctions

강화학습 기반의 온라인 광고 비딩

August 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Sooyoun Park

Efficient Exploration for Online Advertising Auctions

Adviser Min-hwan Oh

Submitting a master's thesis of
Data Science

June 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Sooyoun Park

Confirming the master's thesis written by
Sooyoun Park
July 2023

Chair ____Jaejin Lee____(Seal)

Vice Chair ____Min-hwan Oh____(Seal)

Examiner ____Sanghack Lee____(Seal)

Abstract

This paper addresses the challenges of real-time bidding, where advertisers bid in real-time to win an auction and maximize profit. We propose a novel algorithm that jointly optimizes ad allocation and bidding policy in an online fashion. While previous works have focused on learning either of these components or learning in offline settings with pre-collected data, our algorithm is designed for online advertising auctions. In online advertising auctions, the censored feedback which is provided only when the advertiser wins the auction poses a challenge. Therefore, a proper exploration strategy is essential for learning in online advertising auction environments. Our algorithm integrates exploration in ad allocation and exploration in bid price in an elegant way, using optimistic estimation and count-based control terms. We employ neural networks to estimate the value of each ad and the probability of winning given contextual information. By efficiently collecting data and exploring the dynamic auction environment, our approach outperforms baseline algorithms that do not perform systematic exploration. Additionally, we extend the method to multiple bidding scenarios where agents participate in different auctions. Overall, this paper presents a novel learning algorithm for online advertising auctions which jointly optimizes ad allocation and bidding. We highlight the importance of proper exploration in online advertising auctions as well as the extension to multiple bidding scenarios, by comprehensive experiments.

Keyword: Deep Reinforcement Learning, Real-time Bidding, Exploration
Student Number: 2021-21352

Table of Contents

Chapter 1

INTRODUCTION

Real-time bidding (RTB) [1] has become a prominent framework in the field of online advertising, where ad impressions are sold and bought in real-time. Advertisers participating in auctions aim to maximize user response and profit by strategically determining bid prices for ad impression opportunities. The valuation of ad impressions depends on metrics like conversion rate (CVR), which quantifies user response to ad impressions, such as clicks or sales. Accurate estimation of these valuation parameters is essential for bid price optimization, as over- or under-estimation can lead to suboptimal policies. While contextual bandit algorithms have been widely used for CVR estimation in recommender systems [2, 3], their application to the auction setting has been relatively limited. In contextual bandit algorithms, an agent selects the best possible action based on contextual information, such as user features or auction details. This allows for estimating CVR of ad given the context in auction setting as well.

On the other hand, existing studies on bid optimization primarily assume accurate CVR estimation [4, 5, 6] and overlook the optimization of bidding strategies in the presence of estimation uncertainty. The focus of bid optimization has primarily been on determining optimal bid prices based on the assumption that CVR is known precisely, meaning that we know the optimal ad and its valuation. It is important to note that bid optimization and CVR estimation are interrelated components in the overall advertising process. If the estimation of CVR is inaccurate, it can lead to suboptimal bidding decisions and potentially lower performance. Some prior works have attempted to address the combined optimization of CVR estimation and bid strategies [7, 8]. However, these works have predominantly focused on offline settings where data is readily available. [9] suggested the framework of learning CVR and optimizing bid price while interacting with an auction environment. This paper follows their problem setting. However, they updated the bidding policy every $2000 \sim 10000$ steps and discard all previous data after each update. This training protocol is a limitation of their algorithms when applied to online setting.

We aim to optimize both ad allocation, which involves estimating the CVR of a given ad, and bidding policy, which determines the bid price, in the online setting of advertising auctions. We propose a novel approach that integrates contextual bandit-based exploration within the context of real-time bidding. The challenge of optimizing both ad allocation and bidding policies in an online setting lies in the difficulty of collecting experience and acquiring information about auctions. Advertisers only receive feedback when they win an auction while losing auctions provide no learning opportunities. This scarcity of feedback makes it crucial to devise effective exploration strategies to gather valuable information about the auction environment. We propose Double-UCB, which incorporates optimistic estimation and count-based overbidding. By considering additional optimistic estimations and incorporating count-based overbidding, our approach aims to strike a balance between exploration and exploitation in bid optimization.

To evaluate the effectiveness of our exploration-incorporated bid optimization methods, we conduct extensive experiments in an auction simulation environment. Our results demonstrate that our algorithm achieves smaller regret compared to baseline methods that do not perform systematic exploration. Furthermore, we extend our work to the multi-bidding setting where agents are trained simultaneously to participate in multiple auctions. The exploration from each agent facilitates coordinated exploration as a whole, leading to improved overall performance and effective learning.

In summary, our work makes the following contributions:

- We present an efficient exploration approach for advertising auctions. To our best knowledge, this is the first work to incorporate exploration in bid optimization within both single-bidding and multi-bidding settings.
- We propose an integrated framework optimizing both ad allocation and bidding strategies.
- Through comprehensive experiments employing various methods, we validate the effectiveness of exploration-incorporated bid optimization in improving performance and CVR estimation.

Chapter 2

RELATED WORK

2.1 Bid Optimization in RTB

Budget-constrained bidding is a common strategy in bid optimization. [10] proposed non-linear and concave bidding functions and showed that it is more effective than linear function in a budget-constrained setting. Non-linear functions can encourage higher bids for impressions than their valuation to increase the probability of winning. To address the bid optimization problem, various reinforcement learning approaches have been suggested, such as those proposed in [11, 12, 13]. [12] suggested a model-based reinforcement learning (RL) for learning the optimal bidding policy, where the state is represented by auction-related information including budget and action is the bid price. [13] suggested model-free reinforcement learning, which is more computationally efficient and scalable compared to the model-based method. Another line of research focuses on RTB without budget constraints, specifically maximizing expected revenue. [4, 5, 6] proposed bid shading algorithms to prevent overpayment in first-price auctions. They model the win-rate estimator which estimates the probability of winning as a function of bidding and approximates the optimal bid accordingly. In an adversarial auction environment, [14] proved a regret bound with partial feedback of the dynamics. [9] formulates a bandit-based RTB framework. They split the bid learning process into two steps: ad allocation and bidding. Ad allocation determines which ad should be selected and the bidding part determines the bid price. Previous works often utilize offline datasets and aim to maximize final performance measures. In contrast, our focus is on maximizing online performance during learning, performance measured by regret. We adopt the framework introduced in [9] and simultaneously optimize both allocation and bidding strategies in a coordinated manner. Moreover, our approach is the first to incorporate exploration methods and investigate the impact of exploration on bid optimization.

2.2 Exploration

The Upper Confidence Bound (UCB) algorithm is one of the main contextual bandit algorithms [15, 2, 16, 17, 18] that effectively balances exploration and exploitation. UCB tracks the uncertainty of value estimation of each action and selects the action with the highest upper confidence bound. Bootstrapped UCB [19] leverage bootstrap methods to estimate uncertainty. Contextual bandit algorithms have been applied to recommender systems [2] where state space is user information, action is item, and reward is user response. In environments like bidding, where feedback is often scarce for advertisers, we believe that proper exploration is crucial to gather more information about the auction environment and find an optimal bid policy. We employ the Bootstrapped UCB algorithm in our ad allocation model. This algorithm leverages bootstrap methods to approximate the uncertainty of estimated conversion rates, allowing us to effectively explore different ads based on context information.

2.3 Multi-Agent and Concurrent RL

The field of RL with multiple agents focuses on effective cooperation between agents. A prevalent approach in recent studies is centralized training and decentralized execution (CTDE). An example of this is MADDPG introduced by [20], which demonstrated effective learning in both cooperative and competitive multi-agent environments. Another method that falls within the CTDE framework is COMA proposed by [21]. In the context of real-time bidding (RTB), [8] suggested a multi-agent RL-based algorithm. They presented a coordinated bidding model in which agents cooperate for a common objective. [22] suggested concurrent reinforcement learning, where agents learn concurrently in a common environment. They proposed a seed sampling method, where each agent samples seed independently and maps MDP accordingly. In this paper, we propose a multi-bidding setting where agents are trained simultaneously to participate in multiple auctions. This differs from the conventional multi-agent setting, where agents cooperate within a single environment. Agents share experiences with each other and perform centralized learning. This way, agents can benefit from sharing their experiences and collectively improve their bidding strategies.

Chapter 3

PRELIMINARIES

Optimizing ad allocation and bidding simultaneously in the presence of estimation uncertainty poses significant challenges. Previous studies have typically focused on addressing either ad allocation or bid optimization separately, rather than tackling both together. Some attempts have been made to combine CVR estimation and bid strategies, but these have mostly been confined to offline settings with readily available data or extended learning horizons, which can be considered almost offline. However, these offline approaches do not accurately reflect the real-world scenario faced by new advertisers with limited prior knowledge participating in live auctions. To address these limitations, we propose an integrated ad allocation and bid optimization algorithm designed for online auction environments. Our approach aims to efficiently gather data and gain a better understanding of the dynamic auction environment through effective exploration, ultimately leading to a small regret in bid optimization.

3.1 Problem Setting

We formalize the online advertising auction problem as a contextual bandit problem. At each time step, an advertiser receives a context vector $x \in X$, which encodes the information from the ad impression opportunity. The action of the advertiser consists of two components. Firstly, the advertiser selects one ad i_t from the inventory of ads $[N_i]$. Each ad is described by its feature $a_j \in A (j \in [N_i])$ and its private valuation $v_j \in \mathbb{R}^+$. Secondly, the advertiser bids $b_t \in \mathbb{R}^+$ for the impression opportunity.

Multiple advertisers participate in the auction, and we assume the auction is a first-price auction. Therefore the one with the highest bid takes the impression opportunity and it is charged the price equivalent to its bid. For the advertiser of our interest, we denote the random binary variable indicating a successful bid, w_t, and the binary random variable indicating the occurrence of a conversion event, o_t. Using these notations, the reward at time step t is $r_t = w_t(o_t - b_t)$. We assume that w_t and o_t are conditionally independent given (x_t, i_t, b_t), which means the probability of winning the auction is independent of the probability of conversion events. Finally, we assume that the environment is stationary.

3.2 Learning Objective

The goal of the advertiser is maximizing its expected reward, which is the profit from advertising. This is equivalent to minimizing the (pseudo)regret $\sum_t E_{\pi^*}[r_t] - E_\pi[r_t]$ where π is the policy of the advertiser and π^* is the optimal policy. Here the policy determines which ad to place and how much to bid. Following [9], we call the former the allocation problem and call the latter the bidding problem.

Given a context x, a general stochastic policy $\pi(b, i|x)$ can be factorized as $\pi_b(b \mid i, x)\pi_i(i \mid x)$. Then π_i is the *allocator* and π_b is the *bidder*. The expected reward given the context x is

$$E_\pi[r \mid x] = E_\pi \left[E\left[w(o-b) \mid b, i, x\right] \mid x\right] = E_\pi[E[w \mid b, x]E[o \mid i, x] - E[w \mid b, x]b \mid x] \tag{3.1}$$

With the independence assumption from the problem setting. The conditional expectations $E[w \mid b, x]$ and $E[o \mid i, x]$ are irrelevant to the policy, hence we write $P_w(b, x) = E[w \mid b, x]$ and $P_o(i, x) = E[o \mid i, x]$ to clarify that they are functions of (bid, context) and (ad, context), respectively. Further simplifying the expression, we get

$$\begin{aligned} E_\pi[r \mid x] &= E_\pi[P_w(b, x)(P_o(i, x) - b) \mid x] && (3.2) \\ &= \sum_i \pi_o(i \mid x) \int \pi_b(b \mid i, x) P_w(b, x)(P_o(i, x) - b)\, db && (3.3) \end{aligned}$$

Given an ad i, the inner integral of the last expression of (3.3) is maximized when π_b is the deterministic policy $\pi_b(i, x) = \arg\max_b P_w(b, x)(P_o(i, x) - b)$. Therefore the entire integral is maximized when π_i is also deterministic, $\pi_i(x) = \arg\max_i \max_b P_w(b, x)(P_o(i, x) - b)$. Note that for any context x and any two ads $j_1, j_2 \in [N_i]$, $P_o(j_1, x) \leq P_o(j_2, x)$ implies $\max_b P_w(b, x)(P_o(j_1, x) - b) \leq \max_b P_w(b, x)(P_o(j_2, x) - b)$. Hence we get this intuitive result on optimal policy,

$$i^* \in \arg\max_i P_o(i, x) \tag{3.4}$$

$$b^* \in \arg\max_b P_w(b, x)(P_o(i^*, x) - b) \tag{3.5}$$

The equation (3.5) represents the general term used for estimating win rates in advertising, as discussed in previous works [7, 4, 5]. The advertiser has to select the ad with the highest chance of conversion events, then bid according to (3.5). In other words, learning the optimal policy is optimizing two functions, $P_w(\cdot, \cdot)$ and $P_o(\cdot, \cdot)$.

3.3 Challenges in Advertising Auction Problem

We decompose the advertising auction problem into two parts: the ad-allocation problem and the bidding problem. From the optimal policy (3.4) and (3.5), the former is equivalent to learning a CVR estimator $\hat{P}_o(\cdot, \cdot; \theta_o)$ and the latter can be solved by learning a win rate estimator $\hat{P}_w(\cdot, \cdot; \theta_w)$. At time t, we train function approximators with dataset $\{(x_j, i_j, b_j, w_j)\}_{j=1,2,\cdots,t}$ gathered until t, and approximate $\hat{P}_o(\cdot, \cdot; \theta_o)$ and $\hat{P}_w(\cdot, \cdot; \theta_w)$. The advertiser continues to bid with the trained estimators.

Given a context x, a win rate estimator $\hat{P}_w(\cdot, \cdot; \theta_w)$ and CVR estimation $\hat{P}_o(i, x; \theta_o)$ of the selected ad i, the bidder can determine the bid price using (3.5). For instance, a grid search over bidding value b to maximize $\hat{P}_w(b, x; \theta_w)$ $(\hat{P}_o(i, x; \theta_o) - b)$ could be used. This approach, commonly referred to as bid shading [18, 5] is widely employed in the field.

However, as we will demonstrate later, this strategy may yield subpar performance due to its lack of exploration. In particular, This lack of exploration becomes more critical in advertising auction scenarios, and there are two major challenges.

Censored Feedback. The allocation problem shares a similar objective to contextual bandits, aiming to find an optimal ad given a context such as auction information. However, unlike standard contextual bandit settings, the feedback for action in the allocation problem is censored when the advertiser loses the auction. Whether feedback is received or not depends on the bid and the dynamics of the auction environment. In the presence of censored feedback, the advertiser faces the challenge of balancing the level of overbidding. Overbidding leads to an immediate loss in terms of monetary expenditure, but it also contributes to improving CVR estimations by accelerating data aggregation. This situation presents the bidder with an exploration-exploitation dilemma.

Underbidding. During the early stage of training, when the advertiser has yet to identify the optimal ad, a suboptimal ad may be selected. The estimated CVR of the chosen ad is likely to be lower than that of the optimal ad. As in (3.5), bidding with this estimation leads to underbidding compared to the optimal bidding strategy. Consequently, the advertiser consistently loses auctions, impeding the learning process. We observed this failure mode, where the agent loses almost every auction, even with the perfect win rate estimator.

Chapter 4

PROPOSED METHODS

We propose an integrated algorithm that is capable of learning both the CVR estimator and win rate estimator in an online setting.

4.1 Ad Allocation

The goal of the ad allocation problem is to find the optimal ad given a context. Although it suffers from censored feedback, the training procedure of the allocator does not change; it should learn which ad has the highest CVR given past, uncensored feedback. Therefore we interpret the allocation problem as a contextual bandit and employ the epsilon-greedy algorithm and the UCB algorithm in conjunction with neural network CVR estimators. Also, we introduce a count-based bonus to encourage exploration in the early stage.

Upper Confidence Bound. Upper confidence bound (UCB) [15, 2, 17] algorithms take into account the uncertainty of estimation. The UCB algorithm selects the action with the highest upper confidence bound, which is the sum of the estimation and uncertainty. To estimate the uncertainty of the conversion rate, we employed a bootstrap [23, 19] approach using K estimators. We train neural networks with K heads. The uncertainty is approximated by calculating the standard deviation across the K estimates.

Count-Based Bonus. As we utilize neural networks as function approximators, we have observed that even with the allocation strategies based on UCB, there can be instances where exploration of all available actions is insufficient. To mitigate this initial bias, we introduce a count-based bonus. Count-based bonus is added to the estimated CVR in order to encourage further exploration of different actions. Subsequently, the ad with the highest sum of the original CVR estimate and the count-based bonus is selected. To calculate the count for each ad, we maintain a Gram matrix of the context vectors from which the ad is selected. Specifically, given a context vector x, the count of ad i is computed as $\frac{||x||}{x^T {G^i_{alc}} x}$ multiplied by a tunable parameter. Here, G^i_{alc} represents the Gram matrix of ad i.

4.2 Bidding

In Section 3.3, we highlighted the significance of overbidding to mitigate bid failure. To address this challenge, we suggest an enhanced bidding strategy that integrates uncertainty estimation techniques to effectively induce overbidding.

The advertiser should overbid to collect data to train CVR estimator. Hence, it is desirable for the degree of overbidding to reflect the uncertainty in CVR estimation. To achieve this, we replace the CVR estimation $\hat{P}_o(i, x; \theta_o)$ in 3.5 with *optimistic estimation.* This replacement implicitly encourages overbidding by shifting the location of the maximum to the right. The extent of overbidding can be controlled by adjusting the scale of the uncertainty term. Additionally, we augment the exploration strategy with count-based overbidding. Similar to the count-based bonus employed by the allocator, the bidder maintains a Gram matrix of context vectors from previous steps when the advertiser won the auction.

Double-UCB. We propose a bidding algorithm called Double-UCB, which incorporates optimistic bidding based on the estimated CVR. In our approach, we employ the UCB algorithm to estimate the CVR and scale the overbidding based on the uncertainty associated with the selected ad. The name Double-UCB reflects the optimistic values employed in both the allocation and bidding policies. In Double-UCB, we bid with optimism by taking into account the uncertainty in the CVR estimation. Additionally, we incorporate count-based overbidding, which explicitly promotes exploration in bid prices. As time progresses, the uncertainty in our estimation and the count-based overbidding term both decrease. Consequently, the degree of overbidding gradually decreases over time. There are 4 tunable parameters: the allocation bonus parameter c_{alc} scales the allocation bonus term, thus large c_{alc} generates strong pressure of selecting all ads evenly. The optimism parameter c_{opt} determines the scale of optimistic estimation used in both UCB allocation and bidding. Finally, the overbidding parameter c_{over} determines the degree of count-based overbidding. Algorithm 1 describes the pseudocode for the Double-UCB algorithm. Note that we clip the bid price not to exceed the valuation of the selected ad multiplied by a constant k. In our experiment, k is set to 1.5. Also, we implement the algorithm to update G_{bid} every T_u step.

Initialization of Win Rate Estimator. During our investigations, we observed that the win rate network tends to have a flat shape immediately after initialization. In other words, it initially behaves as a nearly constant function with respect to the bidding value. However, such a flat win rate estimator can impede the learning process, as it exacerbates underbidding by shifting the location of the maximum towards the left. To overcome this challenge, we introduced dummy samples during the initialization of the win rate networks.

Algorithm 1 Double-UCB

```
**Input:** The number of ads $N_i$, Ad features $\{a_j\}_{j\in[N_i]}$, Private valuations $\{v_j\}_{j\in[N_i]}$, An auction environment Auction$(\cdot,\cdot)$, update interval $T_u$, parameters $c_{alc}$, $c_{opt}$, $c_{over}$, $k$
**Initialize:** CVR estimator using UCB $\hat{P}_o(\cdot,\cdot;\theta_o)$, Win rate estimator $\hat{P}_w(\cdot,\cdot;\theta_w)$, $G^j_{alc} = 0$ for all $j \in [N_i]$, $G_{bid} = I$, Memory $\leftarrow \emptyset$
**for** step $t = 1, 2, \cdots$ **do**
    Observe $x_t \in \mathbb{R}^{N_c}$
    $\hat{V}_j, U_j \leftarrow \hat{P}_o(j, x_t; \theta_o)$
    $i_t = \arg\max_j \hat{V}_j + c_{opt}U_j + c_{alc}\frac{||x_t||}{\sqrt{x_t^T G^j_{alc} x_t}}$
    $b' \leftarrow \arg\max_b \hat{P}_w(b, x_t; \theta_w)(\hat{V}_{i_t} + c_{opt}U_{i_t} - b_t)$
    $b \leftarrow \min\{b' + c_{over} v_{i_t}\frac{||x_t||}{\sqrt{x_t^T G_{bid} x_t}}, kv_{i_t}\}$
    $w_t, o_t \leftarrow$Auction$(b_t, x_t)$
    Append $(x_t, i_t, b_t, w_t, o_t)$ to Memory
    $G^{i_t}_{alc} \leftarrow G^{i_t}_{alc} + \frac{x_t x_t^T}{||x_t||^2}$
    $G_{bid} \leftarrow G_{bid} + w_t \frac{x_t x_t^T}{||x_t||^2}$
    **if** $t \equiv 0 \mod T_u$ **then**
        Update $\theta_i$ and $\theta_w$ with Memory
    **end if**
**end for**
```

This adjustment facilitates a more effective exploration of bidding strategies and contributes to a more stable learning process for the win rate network.

4.3 Multi-Bidding

Often, advertisers simultaneously participate in multiple auctions. We extend our method to a multi-bidding setting, where multiple agents participate in different auctions and jointly learn the dynamics of CVR and win rate. In this

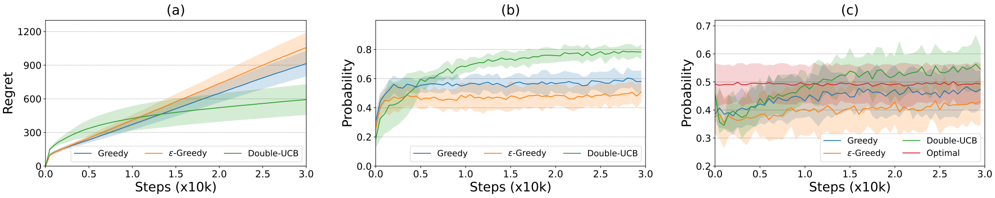


Figure 1: Comparison between Greedy, ϵ-Greedy, and Double-UCB. (a) Regret. (b) Probability of selecting optimal ad. (c) Probability of winning (averaged over 5 runs). The result with the best parameter set $(c_{alc}, c_{opt}, c_{over})$ is reported.

Regret	DM	IPS	DR	Greedy	ϵ-Greedy	Double-UCB (Ours)
$N_c = 5, N_f = 5, N_i = 10$	3175.6 ± 485.9	3834.4 ± 535.6	4760.4 ± 2585.1	941.5 ± 107.1	1049.1 ± 93.6	$\mathbf{592.9 \pm 134.4}$
$N_c = 5, N_f = 5, N_i = 20$	2778.7 ± 196.2	4879.6 ± 4053.4	5953.7 ± 2110.6	1121.4 ± 170.2	1445.4 ± 240.1	$\mathbf{930.0 \pm 60.2}$
$N_c = 10, N_f = 10, N_i = 10$	2741.5 ± 117.9	3110.5 ± 268.5	5271.6 ± 1552.3	1478.7 ± 129.1	1990.5 ± 141.8	$\mathbf{1216.8 \pm 131.3}$

Table 1: Regret over 30K steps (averaged over 5 runs).

setting, each agent interacts with its respective auction environment and collects data, which is stored in a shared replay buffer. We introduce coordinated exploration by assigning different exploration scales to each agent, promoting diverse levels of exploration across multiple auctions. Agents choose actions based on their respective parameters, and all agents share data and undergo centralized training of the allocation and bidding policy.

Chapter 5

EXPERIMENTS AND DISCUSSION

Our experiments are based on AuctionGym [9], a simulation environment for real-time bidding. We modified the training protocol of AuctionGym [9] to enable continuous online learning, by updating the allocator and the bidder every 100 steps and using all data collected until the time step. The context vectors are sampled from N_c dimensional standard Gaussian. Unless otherwise mentioned, $N_c = 5$, $N_f = 5$, and $N_i = 10$ are used. Given a context vector x and an ad feature vector a, the CVR is modeled as $\sigma(\frac{x}{||x||}^T Ma)$ where $\sigma(\cdot)$ is a sigmoid function and $M \in \mathbb{R}^{N_c \times N_f}$. A total of 3 advertisers, including the one being optimized, participate in the auction at each step. We assume that the competing advertisers have the capability to allocate their best ad from their inventory. The bidding behavior of the advertisers is sampled from a normal distribution with a mean of $0.8v_{max}$ and a standard deviation of $(0.1v_{max})^2$, where v_{max} represents the valuation of the best ad. The specific policies adopted by the competing agents, as well as the number of agents, can vary, given that the auction environment remains sufficiently consistent to train the win rate estimator and accumulate winning experience.

Evaluation Metrics. Evaluation of each strategy is based on regret. The regret on each step is computed by the difference between the expected reward (profit) of the optimal policy and the expected reward (profit) of the advertiser's action.

Network Architecture. For our CVR estimator, we employ a network with $N_c + N_f$ input units (concatenation of context and ad features) and 128 hidden units. The Bootstrap-UCB network consists of $K = 5$ heads. As for the win rate estimator, it comprises $N_c + 1$ input units (concatenation of context and bid price) and 15 hidden units, with the bid price input unit skip-connected to the output layer. To ensure stable training, we initialize the win rate networks with dummy samples, as discussed in Section 4.2.

Baseline Methods. As a baseline method, we include Greedy and ϵ-Greedy allocators without an optimistic bidding policy. These methods do not consider uncertainty when estimating bid prices, but they incorporate allocation bonuses and count-based overbidding to collect initial data for training initiation. For ϵ-Greedy, we use $\epsilon = 0.1$. In addition to Greedy and ϵ-Greedy, we compare our Double-UCB algorithm with the bidding algorithm suggested from [9]: Direct Method (DM), Inverse Propensity Score (IPS), and Doubly Robust (DR). DM is a value-based method, and IPS and DR are off-policy policy-gradient methods. We implement these bidding policies in conjunction with our UCB allocator. To align with their methodology, we search over an update interval $T_u \in [2000, 5000]$ for these methods, and vary the entropy regularization coefficient within the range of $[-0.05, 0, 0.05]$ for IPS and DR.

More details on the simulation environment and the training protocol are described in the supplementary material.

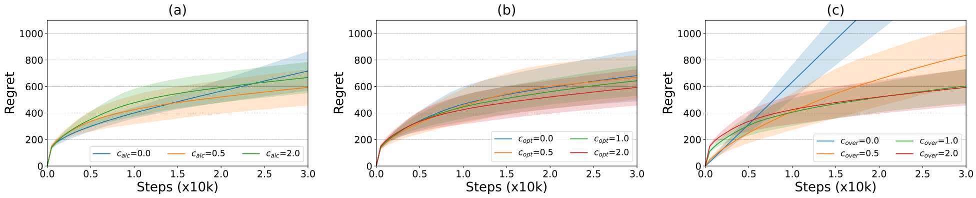


Figure 2: Effect of parameters, (a) c_{alc} (b) c_{opt} (c) c_{over}. Each curve represents the best-performing model with the specified parameter value (averaged over 5 runs).

Parameter	Inclusion / Exclusion							
c_{alc}	×	×	×	○	×	○	○	○
c_{opt}	×	×	○	×	○	×	○	○
c_{over}	×	○	×	×	○	○	×	○
Regret	1906.6	1011.7	1906.6	1905.8	717.0	682.1	1905.8	**592.9**

Table 2: Ablation study of Double-UCB. The ○ mark means the component is used, and the × mark means the component is omitted. For each column, the regret of the best model (over 30K steps, averaged over 5 runs) using ○ marked components is reported.

5.1 Performance Analysis

Table 1 compares the performance of the Double-UCB algorithm with baseline methods. We searched over parameters $c_{alc} \in [0.0, 0.5, 2.0]$, $c_{opt} \in [0.0, 0.5, 1.0, 2.0]$, and $c_{over} \in [0.0, 0.5, 1.0, 2.0]$ for Double-UCB, Greedy, and ϵ-Greedy, then reported the best result. Note that c_{opt} does not apply to Greedy and ϵ-Greedy methods. As expected, Double-UCB outperforms the baselines.

Figure 1 presents the learning curves of Double-UCB, Greedy, and ϵ-Greedy. Both Greedy algorithms initially exhibit smaller regret, indicating rapid exploitation in the early stages. However, their regret steadily increases over time. On the other hand, Double-UCB initially has higher regret due to overbidding, but quickly demonstrates improved performance. This highlights the effectiveness of overbidding in learning ad allocation. For the probability of selecting the optimal ad, Double-UCB consistently outperforms the other methods, while the alternative approaches converge to suboptimal selections. This suggests that Double-UCB provides better estimations of CVR. Comparatively, ϵ-Greedy performs poorly compared to the Greedy method. This indicates that a naive exploration approach is not effective in inducing exploration within the complex and limited-feedback advertising auction environment. It is worth noting that even Greedy and ϵ-Greedy methods outperform DM, IPS, and DR. The incorporation of count-based bonus in Greedy and ϵ-Greedy methods plays a crucial role in initiating learning. This highlights the importance of proper overbidding for successful learning.

5.2 Effect of Parameters

In Figure 2, we analyzed the impact of each parameter, and the results indicate that count-based overbidding has the most significant influence among the three parameters. Insufficient c_{over} can result in learning failure. While allocation bonus and optimistic estimation are not as critical as count-based overbidding, they still enhance performance by promoting exploration in different ways.

Table 2 presents the ablation study, which examines the contribution of each component in our algorithm. The results confirm the previous observation that all three parameters contribute to performance improvement. Specifically, the absence of count-based overbidding leads to failure, further emphasizing its importance in the learning process. It is worth noting that optimistic estimation and count-based overbidding can complement each other, as both encourage overbidding. Optimistic estimation takes into account the uncertainty of CVR estimation, while count-based overbidding provides a more explicit mechanism for overbidding. When count-based overbidding is absent, c_{opt} can play a similar

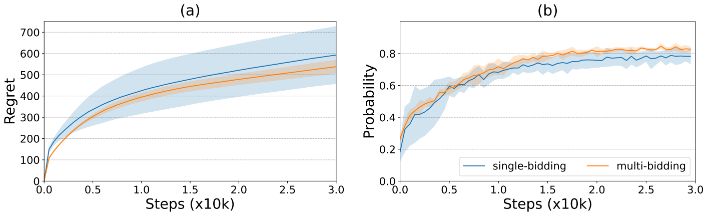


Figure 3: (a) Regret (b) Probability of selecting optimal ad of multi-bidding and single-bidding (averaged over 5 runs). The horizontal axis represents the total number of bidding. Exploration parameters are $c_{alc} = 0.0$, $c_{over} = 1.0$, $c_{opt} \in [0.5, 0.75, 1.0, 1.25, 1.5]$ (uniformly sampled) for each agent.

	Greedy	ϵ-Greedy	Double-UCB (Ours)
Regret	925.4 ± 136.2	1058.9 ± 109.6	$\mathbf{537.9 \pm 30.7}$

Table 3: Regret over 30K steps of Multi-bidding (averaged over 5 runs). The best results for ϵ-Greedy and Greedy are achieved with a constant value of $c_{over} = 1.0$, while the best value for Double-UCB is obtained when c_{opt} is uniformly sampled from the range $[0.5, 0.75, 1.0, 1.25, 1.5]$.

role to c_{over}, but the combination of both parameters yields better performance improvement, as evidenced in Table 2.

5.3 Coordinated Exploration in Multi-Bidding

In a multi-bidding scenario, agents can leverage coordinated exploration by employing diverse exploration scales. Here, each agent possesses a different inventory of ads and participates in different auctions. Through centralized training on allocation and bid policies, agents can share their experiences and collectively improve their performance. In Figure 3, we compare the regret of multi-bidding with 100 agents to single-bidding. The results demonstrate that multi-bidding exhibits smaller regret-per-step and a higher probability of selecting the optimal ad. To further evaluate the performance of different algorithms in the multi-bidding setting, we provide regret values for ϵ-Greedy, Greedy, and Double-UCB in Table 3. Similar to the single-bidding scenario, our proposed method outperforms the baseline. This advantage can be attributed to the two factors. Firstly, the diverse range of ad feature vectors allows for better coverage of the feature space, leading to robust CVR estimation and consequently better selection of the optimal ad. This also results in more consistent behavior across multiple runs, as the experiment demonstrates a smaller variance. Secondly, the accumulation of experiences from multiple agents provides a richer dataset, which enhances the overall learning process. Each agent incorporates a distinct level of optimistic estimation, promoting exploration across different actions based on their respective exploration parameters. For example, agents select one of the parameters from the pool $c_{opt} \in [0.5, 0.75, 1.0, 1.25, 1.5]$, and commit actions with that parameter. This diversity in exploration encourages a broader exploration of the action space and helps discover more optimal strategies. With the centralized training framework, our approach facilitates coordinated exploration and enhances the efficiency of the learning process.

Chapter 6

CONCLUSION

In this paper, we investigate the efficient exploration strategies in the context of online advertising auctions. Our method, Double-UCB, leverages an optimistic estimation of CVR to diversify ad selection and encourage overbidding. Furthermore, Double-UCB employs additional mechanisms: the count-based allocation bonus that prioritizes less selected ads given context, and the count-based overbidding for more robust exploration. We demonstrate that Double-UCB significantly enhances data collection and diversifies the learning experience, resulting in lower regret. Numerical results verify the advantages of Double-UCB over naive exploration methods and other existing bandit-based algorithms, showing reduced regret and improved estimation of CVR. Additionally, our approach proves effective in multi-bidding settings, facilitating robust and efficient learning.

Chapter A

APPENDIX

A.1 Details on Experiments

A.1.1 Training Details

We provide hyperparameters in A.1.1

A.1.2 Implementation Details on Baseline

The original implementation of DM, IPS, and DR from [9] takes a partially offline-fashioned approach. They set large update interval $T_u = [2000, 5000, 10000]$ and data are discarded after each update. To aggregate enough data and initialize the bidding policy, bid prices before the first update are sampled from $N(v, 0.02v)$, where v is the private valuation of the selected ad. This training protocol is to support the importance sampling used in training loss. IPS and DR employ stochastic bidding policies $\pi(\cdot, \cdot; \phi)$ that take contexts and estimated CVR of the selected ads as inputs. For our implementation, we use Bootstrap-UCB as CVR estimators of DM and DR. Following the training protocol of [9], we search over $T_u = [2000, 10000]$ and add entropy regularization with coefficients $[-0.05, 0, 0.05]$. CVR estimators and bidding policy are trained using full batch, 2000 epochs per update. Since this protocol makes

	Hyperparameter	Value
Allocation	Learning rate	1e-3
	Batch size	512
	Update interval	100
	Linear layers	2
	Latent dimension	128
	Bootstrap head	5
Bidding	Learning rate	1e-3
	Batch size	full
	Update interval	100
	Linear layers	2
	Latent dimension	16 (with a skip connection)
Exploration	c_{alc}	0.0, 0.5, 2.0
	c_{opt}	0.0, 0.5, 1.0, 2.0
	c_{over}	0.0, 0.5, 1.0, 2.0

Table A.1.1: Hyperparameters of Double-UCB

CVR estimator unstable to train, we train CVR estimators of DM and DR using the full history, without discarding data.

A.1.3 Training Loss

We describe the training losses of models, given a mini-batch $\{(x_j, i_j, b_j, w_j, o_j)\}$ of size B. For completeness, we include the training losses of DM, IPS, and DR.

CVR Estimator. CVR estimators $\hat{P}_o(\cdot, \cdot; \theta_o)$ only use samples of $w_j = 1$. Then they are trained using binary cross entropy loss:

$$B^{-1} \sum_j o_j \log(\hat{P}_o(i_j, x_j; \theta_o)) + (1 - o_j) \log(1 - \hat{P}_o(i_j, x_j; \theta_o))$$

Win Rate Estimator. Win rate estimators $\hat{P}_w(\cdot, \cdot; \theta_w)$ are also trained using binary cross entropy loss:

$$B^{-1} \sum_j w_j \log(\hat{P}_w(b_j, x_j; \theta_w)) + (1 - w_j) \log(1 - \hat{P}_w(b_j, x_j; \theta_w))$$

DM. DM uses the same loss as Double-UCB. DM employs CVR and win rate estimators as in Double-UCB, but it does not conduct systematic exploration, and uses different training protocol.

IPS. The IPS bidding policy records the probability of the bidding under the current policy at every step. Since they represent the probability of each action under the behavior policy, we compute the importance sampled return:

$$B^{-1} \sum_j \frac{\pi(\hat{P}_o(i_j, x_j; \theta_o), x_j; \phi)}{\pi(\hat{P}_o(i_j, x_j; \theta_o), x_j; \tilde{\phi})} w_j (o_j - b_j)$$

Here gradients do not flow through CVR estimators, and $\pi(\cdot, \cdot; \tilde{\phi})$ is the behavior policy.

DR. DR training loss has additional terms to reduce the variance of the importance sampling term. The term $\hat{P}_o(i_j, x_j; \theta_o)$ is a Monte-Carlo approximation of $\sum_i \pi(\hat{P}_o(i, x_j; \theta_o), x_j; \tilde{\phi}) \hat{P}_o(i, x_j; \theta_o)$.

$$B^{-1} \sum_j \hat{P}_o(i_j, x_j; \theta_o) + \frac{\pi(\hat{P}_o(i_j, x_j; \theta_o), x_j; \phi)}{\pi(\hat{P}_o(i_j, x_j; \theta_o), x_j; \tilde{\phi})} [w_j (o_j - b_j) - \hat{P}_o(i_j, x_j; \theta_o)]$$

A.2 Additional Experimental Results

We present additional experimental results in Figure A.2.1. It shows the performance under various exploration parameters in the single-bidding setting. Figure A.2.2 shows the results of multi-bidding with different exploration hyperparameters.

A.3 Limitations

Our research demonstrates the importance of exploration in learning the policy for advertising auctions in an online setting. The Double-UCB method has proven effective in achieving efficient exploration and maximizing profit in this context. However, our current work does not address budget-constrained auctions, and we leave it to future work to expand our method in a reinforcement learning framework that incorporates the budget as part of the state representation. We expect to develop strategies that we can effectively allocate the budget and maximize overall profit. Additionally, it is desirable to validate our findings by conducting experiments and evaluations in real-world auction settings, thereby assessing the practical applicability and performance of our proposed method.

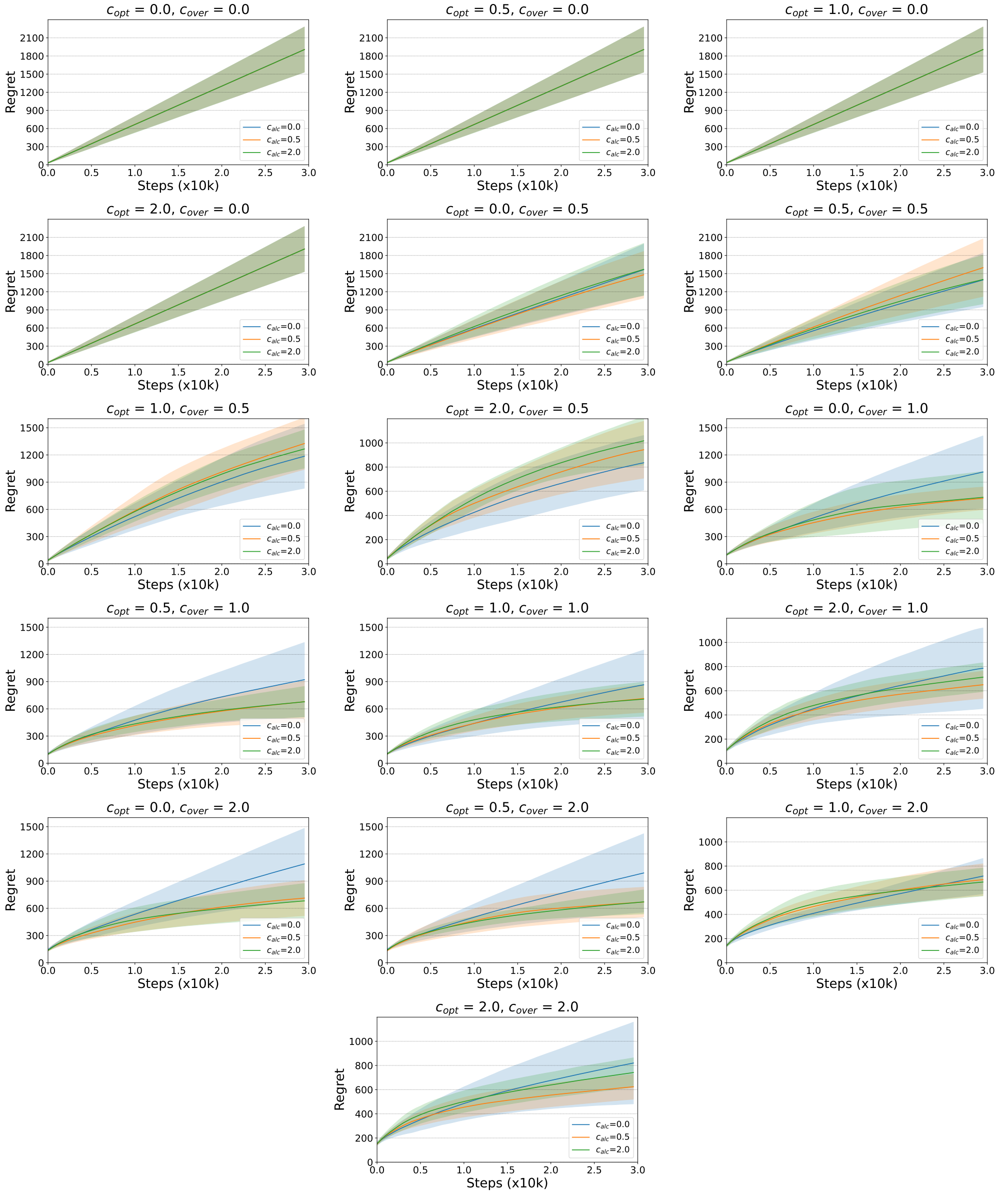
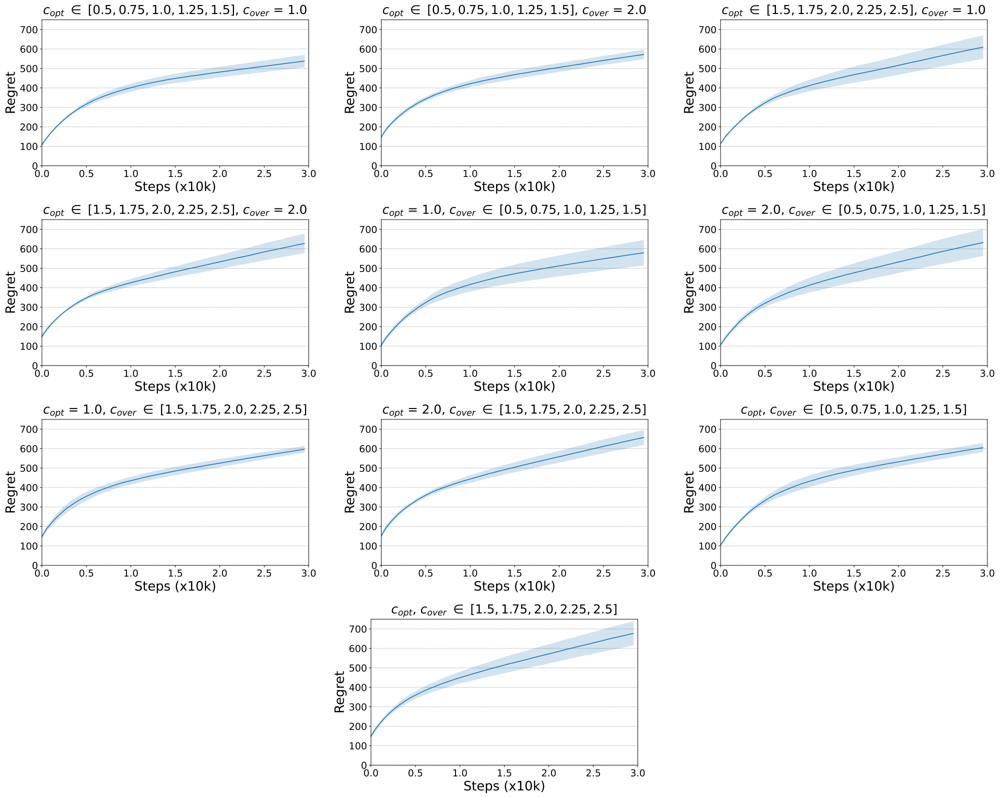


Figure A.2.1: Exploration parameter impact on regret. The result is averaged over five runs.



$c_{opt} \in [0.5, 0.75, 1.0, 1.25, 1.5]$, $c_{over} = 1.0$

$c_{opt} \in [0.5, 0.75, 1.0, 1.25, 1.5]$, $c_{over} = 2.0$

$c_{opt} \in [1.5, 1.75, 2.0, 2.25, 2.5]$, $c_{over} = 1.0$

$c_{opt} \in [1.5, 1.75, 2.0, 2.25, 2.5]$, $c_{over} = 2.0$

$c_{opt} = 1.0$, $c_{over} \in [0.5, 0.75, 1.0, 1.25, 1.5]$

$c_{opt} = 2.0$, $c_{over} \in [0.5, 0.75, 1.0, 1.25, 1.5]$

$c_{opt} = 1.0$, $c_{over} \in [1.5, 1.75, 2.0, 2.25, 2.5]$

$c_{opt} = 2.0$, $c_{over} \in [1.5, 1.75, 2.0, 2.25, 2.5]$

c_{opt}, $c_{over} \in [0.5, 0.75, 1.0, 1.25, 1.5]$

c_{opt}, $c_{over} \in [1.5, 1.75, 2.0, 2.25, 2.5]$

Figure A.2.2: Exploration parameter impact on regret in multi-bidding (100 agents). The result is averaged over five runs. In these experiments, c_{alc} is fixed at 0.0, while c_{opt} and c_{over} are varied within the range of 0.5 to 2.5.

Bibliography

[1] S. Yuan, J. Wang, and X. Zhao, "Real-time bidding for online advertising: Measurement and analysis," 2013.

[2] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

[3] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine," Omnipress, 2010.

[4] D. Gligorijevic, T. Zhou, B. Shetty, B. Kitts, S. Pan, J. Pan, and A. Flores, "Bid shading in the brave new world of first-price auctions," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2453–2460, 2020.

[5] S. Pan, B. Kitts, T. Zhou, H. He, B. Shetty, A. Flores, D. Gligorijevic, J. Pan, T. Mao, S. Gultekin, and J. Zhang, "Bid shading by win-rate estimation and surplus maximization," 2020.

[6] T. Zhou, H. He, S. Pan, N. Karlsson, B. Shetty, B. Kitts, D. Gligorijevic, S. Gultekin, T. Mao, J. Pan, *et al.*, "An efficient deep distribution network for bid shading in first-price auctions," in *Proceedings of the 27th ACM*

SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3996–4004, 2021.

[7] K. Ren, W. Zhang, K. Chang, Y. Rong, Y. Yu, and J. Wang, "Bidding machine: Learning to bid for directly optimizing profits in display advertising," *IEEE Transactions on Knowledge and Data Engineering,* vol. 30, no. 4, pp. 645–659, 2017.

[8] J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. Zhang, "Real-time bidding with multi-agent reinforcement learning in display advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, oct 2018.

[9] O. Jeunen, S. Murphy, and B. Allison, "Learning to bid with auctiongym," 2022.

[10] W. Zhang, S. Yuan, and J. Wang, "Optimal real-time bidding for display advertising," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1077–1086, 2014.

[11] K. Amin, M. Kearns, P. Key, and A. Schwaighofer, "Budget optimization for sponsored search: Censored learning in mdps," 2012.

[12] H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, and D. Guo, "Real-time bidding by reinforcement learning in display advertising," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ACM, feb 2017.

[13] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai, "Budget constrained bidding by model-free reinforcement learning in display advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, oct 2018.

[14] Z. Feng, C. Podimata, and V. Syrgkanis, “Learning to bid without knowing your value,” in *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 505–522, 2018.

[15] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *Journal of Machine Learning Research,* vol. 3, no. Nov, pp. 397–422, 2002.

[16] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” *Advances in neural information processing systems,* vol. 24, 2011.

[17] W. Chu, L. Li, L. Reyzin, and R. Schapire, “Contextual bandits with linear payoff functions,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, JMLR Workshop and Conference Proceedings, 2011.

[18] Z. Zhou, R. Xu, and J. Blanchet, “Learning in generalized linear contextual bandits with stochastic delays,” *Advances in Neural Information Processing Systems,* vol. 32, 2019.

[19] B. Hao, Y. Abbasi-Yadkori, Z. Wen, and G. Cheng, “Bootstrapping upper confidence bound,” 2019.

[20] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” 2020.

[21] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” 2017.

[22] M. Dimakopoulou and B. V. Roy, “Coordinated exploration in concurrent reinforcement learning,” 2018.

[23] B. Efron, *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

초록

강화학습 기반의 온라인 광고 옥션 학습에 대한 논문입니다. 이 논문에서는 광고 할당과 입찰 정책을 공동으로 최적화하는 새로운 알고리즘을 제안합니다. 이전 연구들은 이러한 구성 요소 중 하나를 학습하거나 사전 수집한 데이터와 오프라인 환경에서 학습하는 데 초점을 맞추었으나, 우리의 알고리즘은 온라인 광고 경매를 위해 설계되었습니다. 온라인 광고 경매에서는 학습에 대한 피드백이 경매에서 이길 때에만 제공되기 때문에 학습에 필요한 데이터를 수집하는 것이 어렵습니다. 따라서 온라인 광고 경매 환경에서 적절한 탐험을 통해 학습에 필요한 데이터를 효율적으로 수집하는 것이 중요합니다. 우리의 알고리즘은 Optimistic 추정과 카운트 기반 보너스 사용하여 광고 할당과 입찰 가격에 대한 효율적인 탐험 방식을 제안합니다. 우리는 각 광고의 가치와 정보가 주어졌을 때 경매에서 이길 확률을 추정하기 위해 딥러닝 기반의 모델을 사용합니다. 우리의 알고리즘은 기존의 단순한 탐험 방식과 비교하여 Regret 기준 더 높은 성능을 보였습니다. 또한 우리는 에이전트가 서로 다른 경매에 참여하는 멀티 비딩 환경으로 실험을 확장하여 멀티 비딩에서도 높은 성능을 확인하였습니다. 요약하자면, 이 논문은 광고 할당과 입찰을 공동으로 최적화하는 온라인 광고 경매를 위한 새로운 학습 알고리즘을 제시하고, 온라인 광고 경매에서 적절한 탐험의 중요성을 보여줍니다.

주요어: 강화학습, 온라인 광고 경매, 탐색

학번: 2021-21352