A Unified Solution to Constrained Bidding in Online Display Advertising

Yue He*, Xiujun Chen*, Di Wu*, Junwei Pan§, Qing Tan*, Chuan Yu*, Xu Jian*, Xiaoqiang Zhu*
*Alibaba Group
§Yahoo Research
*{xiaolang.hy,xiujun.cxj,di.wudi,qing.tan,xiaoxun.zhang,yuchuan.yc,xiyu.xj,xiaoqiang.zxq}@alibaba-inc.com
§pandevirus@gmail.com

ABSTRACT

In online display advertising, advertisers usually participate in real-time bidding (RTB) to acquire ad impression opportunities. In most advertising platforms, a typical impression acquiring demand of advertisers is to maximize the sum value of winning impressions under budget and some key performance indicators (KPIs) constraints, (e.g. maximizing clicks with the constraints of budget and cost per click upper bound). The demand can be various in value type (e.g. ad exposure/click), constraint type (e.g. cost per unit value or ad click through rate) and constraint number. Existing works usually focus on a specific demand or hardly achieve the optimum. In this paper, we formulate the demand as a constrained bidding problem, and deduce a unified optimal bidding function on behalf of an advertiser. The optimal bidding function facilitates an advertiser calculating bids for all impressions with only m parameters, where m is the constraint number. However, in real application, it is non-trivial to determine the parameters due to the non-stationary auction environment. We further propose a reinforcement learning (RL) method to dynamically adjust parameters to achieve the optimum, whose converging efficiency is significantly boosted by the recursive optimization property in our formulation. We name the formulation and the RL method, together, as Unified Solution to Constrained Bidding (USCB). Comprehensive experiments on industrial datasets are conducted to demonstrate the effectiveness of USCB. Our solution is deployed and daily impacts millions of revenue in a real-world advertising platform.

KEYWORDS

Real-Time Bidding, Display Advertising, Bid Optimization, Reinforcement Learning

ACM Reference Format:
Yue He*, Xiujun Chen*, Di Wu*, Junwei Pan§, Qing Tan*, Chuan Yu*, Xu Jian*, Xiaoqiang Zhu*. 2018. A Unified Solution to Constrained Bidding in Online Display Advertising. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

KDD '21, August 14–18, 2021, Singapore

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In recent years, online display advertising has become one of the most influential business, with $59.8 billion revenue for FY 2019 in US alone [14]. Real-time bidding (RTB) achieves great success in online display advertising [21, 26], in which, advertisers are required to simultaneously deliver their bids to compete for each ad impression opportunity. The ad impression opportunity would be allocated to the bidder with the highest bid, and the cost is the second highest bid. This allocation and payment rule is called Second Price Auction (SPA) and it plays an important role in online display advertising industry [3, 9].

Nowadays, advertisers participated in the auction of online display advertising can be roughly classified into two categories, one is the brand advertisers and the other is the performance advertisers [22]. Brand advertisers aim for long-term growth and awareness. They construct ad campaigns and typically have the goal of showing their ads to as many audience as possible under some constraints. The constraints are usually related to "shallow" performance indicators such as the average cost per impression/click. On the other hand, the performance advertisers usually try to maximize the sum value of winning impressions and constrain the ad delivery with some "deep" performance indicators, e.g. maximizing conversion number with an average cost per conversion constraint. To satisfy these demands, advertising platforms usually provide corresponding bidding strategies to their customers, such as Google, Facebook, and Alibaba [2, 10, 12].

In order to satisfy different demands of the advertisers, existing works usually focus on solving a specific problem. [23, 25] try to optimize the goal of a campaign under one and two constraints respectively. However, constraints usually vary in both type and number, that is, there could be one or more constraints on different key performance indicators (KPIs), including but not limited to budget, cost per mille (CPM), cost per click (CPC), cost per action (CPA), return on investment (ROI), click through rate (CTR) and Conversion Per Impression(CPI). Besides, although [17] ties to provide a unified solution, the bidding function is not optimal since it aims to minimize the KPI error, which would deliver poor result when KPI constraint is inappropriate, e.g. maximizing clicks with 100$ budget and 100$ CPC constraints[1]. Therefore, it is of great necessity to design a unified solution to optimally satisfy various advertiser demands and be applicable in real-world advertising system.

[1]In real application, there possibly are lots of impressions with lower CPC than 100$, so that the actual CPC in maximizing clicks would be much lower.

In this paper, we abstract the essential demand of advertisers in online display advertising and formulate it as a constrained bidding problem via linear programming [7], i.e. an ad campaign has a core objective of maximizing the sum value of winning impressions under one or more constraints such as budget constraint and key performance indicator constraints. Further, in order to facilitate each ad campaign participating in the auction and enjoying the liquidity of RTB, we leverage the primal-dual method [4] to derive a unified optimal bidding function for the constrained bidding problem. The optimal bidding function is determined by m core parameters, where m is the constraints number.

However, determining the core parameters is not a trivial task. Since the impressions arrive sequentially in a day, it is of great challenge to calculate the core parameters in advance without the complete impression set. Meanwhile, from the perspective of a specific ad campaign, the auction environment is dynamic and unpredictable [23]. As a result, the core optimal parameters from the historical data may be significantly deviating from the actual optimal ones. Thus we intuitively consider it as a sequential parameter adjustment problem based on the basic formulation of constrained bidding. We try to address this problem via reinforcement learning (RL) [16], and the converging efficiency is significantly boosted due to the recursive optimization property of the constrained bidding formulation

To evaluate the effectiveness and generality of our solution to constrained bidding problems, we first construct real-world industrial datasets, based on which, we compare our method with the state-of-the-art methods for three different constrained bidding problems. Experimental results show that our method significantly outperforms other methods.

Our contributions can be summarized into three aspects:

- As far as we know, we are the first to derive a unified optimal bidding function for the constrained bidding problems in online display advertising.
- We propose a reinforcement learning method to search the critical parameter adjustment policy, which is unified, industry-applicable and effective for constrained bidding problems.
- Empirical evaluations on real-world industrial datasets demonstrate the effectiveness of our solution. Beyond the experiments, this solution has been deployed and justified in Taobao display advertising system.

The rest of this paper is organized as follows. Section 2 conducts the formulation of constrained bidding problem and derives the unified optimal bidding function. In section 3, we propose the unified and practical RL approach. Section 4 discusses the experimental results, followed by related work in Section 5. We conclude the paper in Section 6.

2 CONSTRAINED BIDDING

In most advertising platforms, a common demand of impression acquisition is to maximize the sum value of winning impressions under the constraints of budget and various KPI constraints. With the development of online advertising, the demand can be various in value type (e.g. ad exposure/click/conversion), constraint type (e.g. CPC/CTR/ROI) and constraint number. Existing works usually focus on a specific demand [23, 25] or hardly achieve the optimum [17]. In this section, we firstly formulate various demands as constrained bidding problems and then derive a unified optimal bidding function to achieve the optimum.

Table 1: Common KPI constraints in constrained bidding. CTR denotes click through rate, CPI denotes conversion per impression, R denotes payments generated in a conversion and c_i is the cost for winning impression i.

Constraint	$\mathbb{c}_{ij}$	$\mathbb{p}_{ij}$	Type	$\mathbb{k}_j$
CPM	c_i	1×10^{-3}	CR	CPM upper bound.
CPC	c_i	CTR_i	CR	CPC upper bound.
CPA	c_i	CPI_i	CR	CPA upper bound.
ROI	c_i	$\text{CPI}_i \times R$	CR	Reciprocal of ROI lower bound.
CTR	1	CTR_i	NCR	Reciprocal of CTR lower bound.
CPI	1	CPI_i	NCR	Reciprocal of CPI lower bound.

2.1 Problem Formulation

During a time period, one day for instance, suppose there are N impression opportunities arriving sequentially and indexed by i. In a RTB system with SPA, advertisers submit bids to compete for each impression in real-time. An ad campaign[2] will win impression i if its bid b_i is greater than the highest bid c_i of other campaigns. The cost of the impression opportunity is also c_i.

During the impression acquisition, a common goal of an ad campaign is to maximize the sum value of winning impressions, i.e., $\max \Sigma_i v_i x_i$, where v_i is the impression value and x_i is the binary indicator of whether the campaign wins impression i. Besides, budget and KPI constraints are critical for an ad campaign to control the ad delivery performance. Budget constraint can be considered as $\Sigma_i c_i x_i \leq B$. KPI constraints are more complicated and can be classified into two categories. The first category is the cost-related (CR) constraints, which restricts the unit cost of certain advertising event, such as CPC and CPA. The second category is the non-cost-related (NCR) constraints, which restricts the average advertising effect, such as CTR and CPI. The unified expression of KPI constraint indexed by j can be expressed by Eq. (1)

$$\frac{\Sigma_i \mathbb{c}_{ij} x_i}{\Sigma_i \mathbb{p}_{ij} x_i} \leq \mathbb{k}_j \tag{1}$$

, where $\mathbb{k}_j$ is the upper bound of constraint j which is provided by the advertiser, $\mathbb{p}_{ij}$ can be any performance indicator or constant. $\mathbb{c}_{ij} = c_i \mathbb{1}_{CR_j} + \mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j})$, where $\mathbb{q}_{ij}$ can be any performance indicator or constant and $\mathbb{1}_{CR_j}$ is the indicator function of whether constraint j is CR. To facilitate a better understanding, we list some common types of KPI constraints in Tbl. 1.

In summary, considering the advertising goal, budget and M KPI constraints, the common demand of an ad campaign can be formulated as a unified constrained bidding problem by (LP1).

[2]In real-world application, an advertiser usually constructs multiple ad campaigns to compete for impressions.

$$\begin{aligned} \underset{x_i}{\text{maximize}} \quad & \Sigma_i v_i x_i \\ \text{s.t.} \quad & \Sigma_i c_i x_i \le B \\ & \frac{\Sigma_i \mathbb{c}_{ij} x_i}{\Sigma_i \mathbb{p}_{ij} x_i} \le \mathbb{k}_j \quad , \forall j \\ & x_i \le 1 \quad , \forall i \\ & x_i \ge 0 \quad , \forall i \end{aligned} \tag{LP1}$$

2.2 Optimal Bidding Strategy

Theoretically, we can obtain the optimal x_i to select the winning impressions by solving the linear programming problem (LP1). In order to obtain (LP1), impressions set are required to be completely known. However, in real-world application the complete impressions set is difficult to obtain since each impression i arrives from a stream and it is impractical to accurately predict for the incoming impressions. Therefore the remaining problem is how to make an online decision of whether to win each impression i or not. It is non-trivial to solve this problem since the number of impressions is usually enormous and the performance of impressions are hardly predictable before arrival. Fortunately, we succeeded in deriving a unified optimal bidding function for this problem, which significantly reduce the dimension of solution space to the number of constraints, as shown in Thm. 2.1.

Theorem 2.1. *The optimal bid for each impression i is as follows, which results in the optimal decision x_i in* (LP1).

$$b_i^* = w_0^* v_i - \Sigma_j w_j^* (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij}) \tag{2}$$

, where $w_0^* > 0$, $w_j^* \in [0, 1]$ if constraint j is cost-related otherwise $w_j^* \ge 0$, $\forall j \in [1, ..., M]$.

Proof. The dual problem of (LP1) is as follows

$$\begin{aligned} \underset{\alpha, \beta_j, r_j}{\text{minimize}} \quad & B\alpha + \Sigma_i r_i \\ \text{s.t.} \quad & c_i \alpha + \Sigma_j (\mathbb{c}_{ij} - \mathbb{k}_j \mathbb{p}_{ij}) \beta_j + r_i \ge v_i, \quad \forall i, \\ & \alpha \ge 0, \\ & \beta_j \ge 0, \quad \forall j, \\ & r_i \ge 0, \quad \forall i \end{aligned} \tag{LP2}$$

, where α, β_j and r_i are dual variables.

According to the definition of $\mathbb{c}_{ij}$, we can rewrite the first inequality in (LP2) as follows

$$\underbrace{(v_i - \Sigma_j \beta_j (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij}))}_{P_{NCR}} - \underbrace{(\alpha + \Sigma_j \beta_j \mathbb{1}_{CR_j})}_{P_{CR}} c_i - r_i \le 0 \tag{3}$$

, where P_{NCR} represents the non-cost-related coefficients and P_{CR} represents the cost-related coefficients.

Let the optimal solution to the primal problem (LP1) be x_i^* and the optimal solution of the dual problem (LP2) be α^*, r_i^* and β_j^*. According the theorem of complementary slackness [8], we can obtain[3]:

$$x_i^* (P_{NCR}^* - P_{CR}^* c_i - r_i^*) = 0, \ \forall i \tag{4}$$

$$(x_i^* - 1) r_i^* = 0, \ \forall i \tag{5}$$

We delicately set the bid of impression i as $b_i^* = P_{NCR}^* / P_{CR}^*$, then we can transform InEq. (3) to InEq. (6) and Eq.(4) to Eq.(7) respectively.

$$(b_i^* - c_i) P_{CR}^* - r_i^* \le 0, \ \forall i \tag{6}$$

$$x_i^* [(b_i^* - c_i) P_{CR}^* - r_i^*] = 0, \ \forall i \tag{7}$$

It can be inferred that,

- If an ad campaign wins impression i, which means $x_i^* > 0$, then according to Eq.(7), $(b_i^* - c_i) P_{CR}^* - r_i^* = 0$. Meanwhile, since $r_i^* \ge 0$, $P_{CR}^* \ge 0$, it can be inferred that $b_i^* \ge c_i$.
- If an ad campaign loses impression i, which means $x_i^* = 0$, we can deduce from Eq.(5) that $r_i^* = 0$. Since $P_{CR}^* \ge 0$, according to InEq. (6), we can obtain $b_i^* \le c_i$.

To sum up, for any impression i, bidding with b_i^* will result in x_i^*, which is the optimal solution of (LP1). Therefore the optimal bid is b_i^* and to be more concise we organize the expression of b_i^* as follows and finally complete the proof of Thm. 2.1.

$$b_i^* = \frac{P_{NCR}^*}{P_{CR}^*} = \frac{v_i - \Sigma_j \beta_j^* (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij})}{\alpha^* + \Sigma_j \beta_j^* \mathbb{1}_{CR_j}} \tag{8}$$

$$= \underbrace{\left(\frac{1}{\alpha^* + \Sigma_j \beta_j^* \mathbb{1}_{CR_j}}\right)}_{w_0^*} v_i \tag{9}$$

$$- \Sigma_j \underbrace{\left(\frac{\beta_j^*}{\alpha^* + \Sigma_j \beta_j^* \mathbb{1}_{CR_j}}\right)}_{w_j^*} (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij}) \tag{10}$$

$$= w_0^* v_i - \Sigma_j w_j^* (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij}) \tag{11}$$

□

According to Thm. 2.1, for a campaign with M constraints and wishes to maximize the sum value of the winning impressions under constraints, the optimal bid as shown in Eq. (2) is determined by $M + 1$ core parameters w_k^*, $k \in [0, ..., M]$. With this bidding function, we can deal with the constrained bidding problem by learning for the optimal $M + 1$ parameters rather than searching for the optimal decision (or optimal bid) for each impression individually. It is worth noting that the number of parameters is usually much smaller than that of impressions, which significantly simplifies the problem. When the candidate impression set is complete, the optimal parameters can be calculated by solving (LP2), and the optimal bidding function defined in Eq. (2) leads to the same optimum of (LP1). More importantly, when the candidate impression set is incomplete the optimal bidding function enables us to approach to the optimum by adjusting the parameters and facilitates the online decision of whether to win each impression or not. Revisiting previous works, [23, 27, 28] deliver an optimal

[3] P_{NCR}^* and P_{CR}^* are P_{NCR} and P_{CR} calculated with α^*, r_i^* and β_j^* respectively.

bidding function $b_i = w_0\ v_i$ to maximize the sum value of winning impressions under a budget constraint, which is special case of $w_k = 0,\ \forall k \in [1, ..., M]$; while [25] proposes an optimal bidding function $b_i = w_0\ v_i + w_1\ \mathbb{k}_{CPC}\ \text{CTR}_i$ to maximize conversion number under budget and CPC constraints, which is the special case of $v_i = \text{CPI}_i$, $\mathbb{p}_{i1} = \text{CTR}_i$, $\mathbb{k}_1 = \mathbb{k}_{CPC}$ and $\mathbb{1}_{CR_j} = 1$.

3 METHOD

Recall that the optimal constrained bidding strategy is defined by Eq. (2), therefore the remaining challenge is to calculate the optimal parameters $w_k^*,\ k \in [0, ..., M]$. Since the traffic pattern will change between different days and the impression set is not complete until the end of the day, it is impossible to obtain those optimal parameters by solving (LP2). Thus, in application, with initial w_k calculated based on historical impression data, it is of necessity to develop a parameter adjustment agent, which applies its policy to modify w_k to the optimal one(s) under the current state (i.e. advertising status of a campaign, including budget spend status, KPI constraint satisfying status, etc.). In this section, we first formulate this problem as an Markov Decision Process, then we introduce an important property of constrained bidding problem, last we present an efficient policy search method via reinforcement learning [20].

3.1 Modeling

We formulate the parameter adjustment problem as a Markov Decision Process. A Markov Decision Process is defined by a set of states $\mathcal{S}$ describing the advertising status of a campaign, a parameters adjustment action space $\mathcal{A} = \mathcal{A}_0 \times \mathcal{A}_1 \times ... \times \mathcal{A}_M \subseteq \mathbb{R}^{M+1}$ for an agent. At each time-step t, an agent delivers actions $a_{0t}, a_{1t}, ..., a_{Mt} \in \mathcal{A}$ based on the current state $s_t \in \mathcal{S}$ to modify $w_{kt}, k \in [0, ..., M]$, according to its policy $\pi : \mathcal{S} \mapsto \mathcal{A}$. Then, the state transitions to a next state according to the state transition dynamics $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \Omega(\mathcal{S})$ where $\Omega(\mathcal{S})$ is a collection of probability distributions over $\mathcal{S}$. The environment returns an immediate reward to the agent based on a function of current state and agent actions as $r_t : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{R} \subseteq \mathbb{R}$. The goal of the agent is to maximize its the total expected return $R = \sum_{t=1}^{T} \gamma^{t-1} r_t$ where γ is a discount factor and T is the time horizon. The detailed description of our modeling are listed below:

- $\mathcal{S}$: The state is a collection of information which describes the advertising status from the perspective of a campaign. The information should in principle reflect time, budget consumption and KPI constraints satisfying status, such as left time, left budget, budget consumption speed, current KPI ratio for constraint j (i.e. $(\Sigma_i \mathbb{c}_{ij} x_i / \Sigma_i \mathbb{p}_{ij} x_i) / \mathbb{k}_j$), etc.
- $\mathcal{A}$: At time-step t, each agent delivers an $M+1$ dimensional action vector $\vec{a}_t = (a_{0t}, ..., a_{Mt})$ to modify the $M+1$ dimensional parameter vector $\vec{w}_t = (w_{0t}, ..., w_{Mt})$, typically taking the form of $\vec{w}_{t+1} = \vec{w}_t \cdot (1 + \vec{a}_t)$, where $a_{kt} \in (-1.0, +\infty),\ \forall k \in [0, ..., M]$.
- r_t : At time-step t, let O be the candidate impression set between step t and $t+1$, then the reward $r_t = \Sigma_{i \in O} x_i v_i$, i.e. the sum value of the winning impressions in O.
- $\mathcal{T}$: We apply a model-free RL method to solve our problem, so the transition dynamics is not required to be explicitly modeled.
- γ : We set reward discount factor $\gamma = 1$ since the effectiveness of each campaign needs to be evaluated from a daily perspective.

3.2 Unified Solution to Constrained Bidding (USCB)

Recall that we have derived the optimal bidding function (2) for the constrained bidding problem. Since the impressions arrive sequentially in a day, at each time step, we will face a recursive optimization problem, which is to maximize the sum value of winning impressions over the remaining impressions. We prove that it can be formulated in the same form as (LP1) and the optimal action for the agent is shown by the following Thm. 3.1.

Theorem 3.1. *For a sub-problem at each time-step t, the optimal action sequence for the bidding agent is to modify the current $\vec{w}_t$ to the optimal $\vec{w}_t^*$, and keep it fixed at the following time steps.*

Proof. At any time-step t, since a campaign has already won several impressions, the sub-problem is how to select from the arriving impressions to optimize the global problem in (LP1) given the winning impressions. In the sub-problem, the impression set, objective and constraints should be refreshed. Specifically, let the value, cost and the performance for impressions that already won are $v^\#$, $c^\#$, $\mathbb{c}_j^\#$ and $\mathbb{p}_j^\#$, $j \in [1, ..., M]$ respectively and the remaining impression set be O. Then the sub-problem can be formulated as follows:

$$\begin{aligned} \underset{x_i, i \in O}{\text{maximize}} \quad & \Sigma_i v_i x_i + v^\# \\ \text{s.t.} \quad & \Sigma_i c_i x_i + c^\# \le B \\ & \frac{\Sigma_i \mathbb{c}_{ij} x_i + \mathbb{c}_j^\#}{\Sigma_i \mathbb{p}_{ij} x_i + \mathbb{p}_j^\#} \le \mathbb{k}_j, \quad \forall j, \\ & x_i \le 1, \quad \forall i \in O, \\ & x_i \ge 0, \quad \forall i \in O \end{aligned} \tag{LP3}$$

The dual problem of (LP3) is as follows:

$$\begin{aligned} \underset{\alpha, \beta_j, r_j}{\text{minimize}} \quad & (B - c^\#)\alpha + \Sigma_j (\mathbb{p}_j^\# \mathbb{k}_j - \mathbb{c}_j^\#)\beta_j + \Sigma_i r_i \\ \text{s.t.} \quad & c_i \alpha + \Sigma_j (\mathbb{c}_{ij} - \mathbb{k}_j \mathbb{p}_{ij})\beta_j + r_i \ge v_i, \quad \forall i, \\ & \alpha \ge 0, \\ & \beta_j \ge 0, \quad \forall j, \\ & r_i \ge 0, \quad \forall i \end{aligned} \tag{LP4}$$

, where α, β_j and r_i are dual variables.

Similar to Section 2.2, we can derive that the optimal bid b_{it}^* for the sub-problem at step t takes the same form as Eq. (2), i.e.,

$$b_{it}^* = w_{0t}^* v_i - \Sigma_j w_{jt}^* (\mathbb{q}_{ij}(1 - \mathbb{1}_{CR_j}) - \mathbb{k}_j \mathbb{p}_{ij}) \tag{12}$$

, where $\vec{w}_t^*$ is the optimal parameter of sub-problem at time-step t. Thus, at any time-step t, the optimal action is to adjust the parameter to $\vec{w}_t^*$. Assume the Markov Decision Process is deterministic,

after taking the optimal action, the state will transition to a subsequent state where the optimal action is to keep the parameter fixed. Therefore, at any time-step the optimal action sequence is to adjust current parameter to $\vec{w}_t^*$ and do not adjust them for the following time-steps. □

We present our method named Unified Solution to Constrained Bidding (USCB) in Algo. 1. Firstly, without loss of generality, we apply DDPG [18] as the implementation of our method. Secondly, also most importantly, based on Thm. 3.1, the learning processes of actor and critic are simplified, which can be interpreted from the following two aspects:

- At time-step t, the agent tries to settle a matter at one go, which is to adjust current $\vec{w}_t$ to $\vec{w}_t^*$ and keep it fixed until time-step T rather than sequentially adjusting it for the rest of the steps. This significantly reduces the learning difficulty of the RL model.
- The learning process of the critic Q is simplified as minimizing the difference between $\mathcal{G}$ and $Q(s_t, \vec{a}_t)$, where $\mathcal{G}$ is a direct signal of differentiating which action would lead to a better result of problem (LP3), and $Q(s_t, \vec{a}_t)$ is the state action value function in RL which indicates the expected accumulated value while taking action $\vec{a}_t$ under state s_t. Compared with common practice of updating $Q(s_t, \vec{a}_t)$ towards $r_t + \gamma Q(s_{t+1}, \pi(s_{t+1}))$ (more details see temporal-difference method [20]), the learning process of Q is easier and, in turn, Q will boost the policy convergence.

4 EXPERIMENTS

In this section, we introduce the experimental setup and report the evaluation results. Finally, we briefly describe the online application of our method. We conduct our empirical study based on three different constrained bidding problems. For each problem, we compare our method with the corresponding state-of-the-art method and show the superiority of our method. The superiority is not only reflected in the critical metric lift, but also in the converging efficiency and rationality. We conduct experiments on the top three constrained bidding products preferred by advertisers in Taobao advertising system:

- CB{click}: The objective is to maximize clicks and the constraint is budget (i.e. $\Sigma_i c_i x_i \leq B$), which is problem solved in [23]. The optimal bidding function for impression i is

$$b_i = w_0 \ \text{pCTR}_i \tag{13}$$

- CB{click-CPC}: The objective is to maximize clicks and the constraints are budget and CPC (i.e. $\Sigma_i c_i x_i / \Sigma_i \text{pCTR}_i x_i \leq \Bbbk_{CPC}$). The optimal bidding function for impression i is

$$b_i = w_0 \ \text{pCTR}_i + w_1 \ \Bbbk_{CPC} \ \text{pCTR}_i \tag{14}$$

- CB{conversion-CPC}: The objective is to maximize conversion number and the constraints are budget and CPC, which is the problem solved in [25]. The optimal bidding function for impression i is

$$b_i = w_0 \ \text{pCVR}_i + w_1 \ \Bbbk_{CPC} \ \text{pCTR}_i \tag{15}$$

Algorithm 1: Unified Solution to Constrained Bidding

Initialize a $M+1$ dimensional random process $\mathcal{E}$;
Initialize replay memory $\mathcal{M}$ with capacity N;
Initialize actor π_θ with weights θ;
Initialize critic Q_η with weights η;
Set Batch Size as BS;
Let $\vec{w}^*$ be the optimal parameter vector $(w_0^*, ..., w_M^*)$;
Let R^* be the theoretically optimal result;
while *not convergent* **do**
  Randomly choose an ad campaign and simulate SPA;
  Set $\vec{w}_1 = \vec{w}^* + \mathcal{E}$;
  Bid with $\vec{w}_1$ at time-step 1 and get reward r_1;
  Set $R = r_1$;
  for $t = 2$ **to** T **do**
    Observe state s_t;
    Get action vector $\vec{a}_t = \pi_\theta(s_t) + \mathcal{E}$;
    Set $\vec{w}_t = \vec{w}_{t-1} \cdot (1 + \vec{a}_t)$;
    Bid with $\vec{w}_t$ at time-step t and get reward r_t;
    Set $R = R + r_t$ and $V = 0$;
    for $\tau = t+1$ **to** T **do**
      Bid with $\vec{w}_t$ at time-step τ and get reward r_τ;
      Set $V = V + r_\tau$;
    end
    Calculate penalty p_j for KPI constraint j;
    Set $\mathcal{G} = \min\{(R+V)/R^*, 1.0\} - \Sigma_j p_j$;
    Store $(s_t, \vec{a}_t, \mathcal{G})$ in $\mathcal{M}$;
    Sample BS $(s^k, \vec{a}^k, \mathcal{G}^k)$ tuples from $\mathcal{M}$;
    Update Critic Q_η by minimizing the loss $\mathcal{L}(\eta) = \frac{1}{\text{BS}} \sum_k (\mathcal{G}^k - Q_\eta(s^k, \vec{a}^k))^2$;
    Update Actor π_θ by policy gradient: $\nabla_\theta J \approx \frac{1}{\text{BS}} \sum_k \nabla_\theta \pi(s^k) \nabla_{\vec{a}} Q^\pi(s^k, \vec{a})|_{\vec{a}=\pi(s^k)}$;
  end
end

Table 2: The statistics of the datasets for three different constrained bidding problems.

Dataset	Campaign No.	Impression No.	Avg. Deviation
CB{click}	3900	3.04 B	10.06%
CB{click-CPC}	3900	4.56 B	13.87%
CB{conversion-CPC}	3974	2.26 B	15.11%

4.1 Experimental Setup

4.1.1 Datasets. We use the real-world bidding log from Taobao display advertising platform to separately build our datasets for the three different constrained bidding problems. Specifically, for each dataset $\mathcal{D}$, each instance can be denoted by $(t_i, cid, \text{pCTR}_i, \text{pCVR}_i, c_i)$, where $t_i, cid, \text{pCTR}_i, \text{pCVR}_i, c_i$ are timestamp, ID of campaign, pCTR, pCVR and winning price of campaign cid for impression i respectively. As shown by Tbl. 2, each dataset contains more than 3900

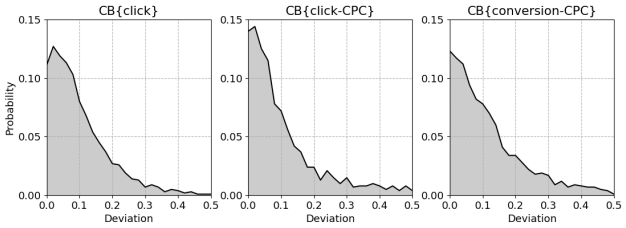

Figure 1: The deviation distributions of the datasets for three different constrained bidding problems.

campaigns and more than 2 billion impressions from 11st Jan. to 23rd Jan. 2021. We use a 8-days time window to determine the training, validation and testing dataset. 90% of the previous 7-days of data, which is randomly sampled, is used for training and remaining 10% is used for validation. The last day of data is used for testing. Therefore, there are 6 sets of training/validation/testing datasets.

Recall that for any optimal bidding function, initial parameters $w_k,\ \forall k \in [0, ..., M]$ should be determined beforehand. In our scenario, for any ad campaign, we use the optimal parameters from the 7th day of training dataset as the initial parameters of testing dataset. We use Eq. (16) to quantitatively describe the difference between the initial parameters and the actual optimal ones. Since $w_0,\ w_0^* \in \mathbb{R}^+$, $w_k \in [0, 1]\ \forall k \in [1, ..., M]$ in our datasets and we ensure the first term in numerator lie in $[0, 1]$, the codomain of Eq. (16) will also lie in $[0, 1]$. The statistics and visualizable distribution of deviations are shown by the last column of Tbl. 2 and Fig. 1. It can be seen that, from the perspective of a campaign, the difference between the initial parameters and optimal ones could be very large, which is caused by the unstable auction environment.

$$Deviation = \begin{cases} \sqrt{\frac{(w_0/w_0^*-1)^2+\Sigma_j(w_j^*-w_j)^2}{M+1}} & \text{if } w_0 \le w_0^*, \\ \sqrt{\frac{(w_0^*/w_0-1)^2+\Sigma_j(w_j^*-w_j)^2}{M+1}} & \text{otherwise} \end{cases} \tag{16}$$

, where, $\forall k \in [0, ..., M]$, w_k is the optimal parameter from the 7th day of training dataset and w_k^* is the optimal one in testing dataset.

4.1.2 Implementation Details. We take a fully connected neural network with 3 hidden layers and 50 nodes for each layer to implement the actor and critic network for DDPG [18]. The batch size is set to 128 and the replay memory size is set to 100,000. Agent outputs continuous actions for each bidding parameter with range of $[-0.2, 0.2]$. The agent takes action in every 15 minutes, so the time horizon T in Algo. 1 is 96. For the exploration random process, we set $\mu = 0$ and $\sigma = 0.05$. The learning rate of actor is set to $1e^{-6}$ and critic is set to $1e^{-7}$. The hyper-parameter λ of KPI constraint violation penalty defined in (17) is set to 20. In our case, the best model selection criterion is choosing the best one (metric would be introduced in next section) on validation dataset within 2000 episodes. In validation and testing process, for each campaign, the initial bidding parameters is set to the optimal parameters of the previous day of data. The optimal parameters can be calculated by GNU Linear Programming Kit (GLPK) to solve the dual problem of (LP1).

4.1.3 Evaluation Metrics. For any ad campaign, the goal of constrained bidding is to maximize the sum value of winning impressions under some constraints. Let $R = \Sigma_{t=1}^{T} r_t$ be the return of the applied policy π and R^* be the theoretically optimal result for the constrained bidding problem. In principle, the metric should consider the following two aspects:

- There should be a penalty term for constraint violation.
- Only the optimal policy could achieve the maximal value.

Thus, we design the metric as Eq. (17). The penalty term (i.e. $\Sigma_j p_j$) would punish the policy if the constraints are violated. The operator "min" along with the penalty term ensure that only the optimal policy could reach the maximum value of 1.0.

$$G = \min\{\frac{R}{R^*}, 1.0\} - \Sigma_j p_j \tag{17}$$

, where $p_j = \lambda^{exr_j} - 1$ is the penalty term for violating constraint j, $\lambda \in (1, +\infty)$ is the penalty strength hyper-parameter, $\mathbb{k}_j^{actual} = \Sigma_i \mathbb{c}_{ij} x_i / \Sigma_i \mathbb{p}_{ij} x_i$, $exr_j = \max\{0,\ \mathbb{k}_j^{actual}/\mathbb{k}_j - 1\}$ indicates the extent of breaking KPI constraint j.

4.1.4 Compared Methods.

- Fixed Parameter (FP): a method that bidding with Eq. (2) and keeping the parameters fixed with the optimal ones of historical data in testing stage.
- Deep Reinforcement Learning to Bid (DRLB): the method delivered in [23] is the state-of-the-art algorithm for single parameter control, which aims for maximizing impression value under a single budget constraint. DRLB is a method based on RL, applying reward shaping technique to boost the convergence.
- Model predictive PID (M-PID): the method proposed in [25] is the state-of-the-art algorithm for double parameters control, which aims for maximizing conversion number under budget and CPC constraints. M-PID is based on PID controller and needs expert domain knowledge to determine control target.
- Unified Solution to Constrained Bidding (USCB): the unified solution to constrained bidding proposed in this paper, which abstracts the core demand of constrained bidding and augments it with a more efficient policy search method.

4.2 Evaluation Results

We conduct experiments to compare the performance of FP, M-PID, DRLB and USCB. As mentioned above, due to the unstable traffic in real-world application, the impression volume and distribution in the testing data deviate from that of training data, which increases challenge of bidding problems. To investigate the performance of each method with different deviations, we divide all campaigns into 3 groups according to the deviation and evaluate the four methods in each group.

The performance of these methods are summarized in Tbl. 3. Our method significantly outperforms FP, M-PID, DRLB on all datasets that correspond to 3 types of constrained bidding problems. In addition, as the deviation increases, the performance of all methods degrade. Our method maintains the superiority on all deviations.

Table 3: The comparison of G of FP, M-PID, DRLB and our method in 3 datasets with 3 levels of deviation.

	CB{click}					CB{click-CPC}					CB{conversion-CPC}				
Deviation	Campaign No.	FP	M-PID	DRLB	USCB	Campaign No.	FP	M-PID	DRLB	USCB	Campaign No.	FP	M-PID	DRLB	USCB
[0.00,0.05]	637	0.962	0.895	0.934	**0.982**	575	0.936	0.917	0.930	**0.983**	543	0.953	0.906	0.927	**0.976**
[0.05,0.20]	939	0.778	0.875	0.912	**0.979**	797	0.891	0.895	0.911	**0.973**	846	0.832	0.883	0.914	**0.962**
[0.20,1.00]	224	0.567	0.867	0.893	**0.966**	428	0.648	0.872	0.905	**0.955**	444	0.541	0.852	0.906	**0.943**
Overall	-	0.817	0.881	0.917	**0.978**	-	0.848	0.897	0.916	**0.972**	-	0.797	0.882	0.916	**0.962**

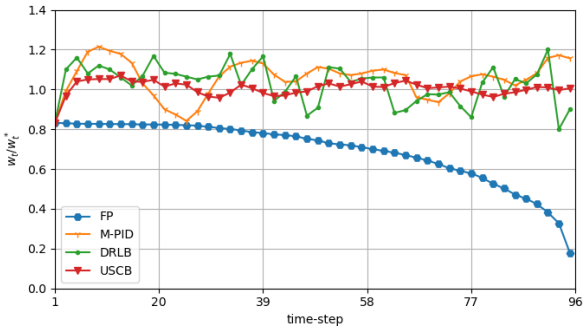

Figure 2: An example of CB{click} bidding parameters adjustment process of all methods. Each method starts with the same parameter. At each step t, it obtains its own parameters $\vec{w}_t$ and the optimal parameters $\vec{w}_t{}^*$ for the current subproblem. The $\vec{w}_t/\vec{w}_t{}^*$ at each step is plotted, which illustrates how close the current parameter is to the optimal parameter. Generally speaking, the closer the $\vec{w}_t/\vec{w}_t{}^*$ to 1.0, the better the result is.

To better understand the performances of all methods, we illustrate an example of the CB{click} bidding parameters adjustment process of all methods in Fig. 2. Based on these results we present the analyses on these methods as follows.

- FP: FP is simple, easy to implement, and could achieve good results when the auction environment is stable. However, since auction environment may fluctuates enormously in the real-world application, bidding with the historical optimal parameters cannot guarantee a satisfactory result. As shown by Tbl. 3, FP delivers the worst overall result due to its lack of adaptability to environment change. Specially, it obtains the second best results when deviations are small, but obtains the worst results as deviations increasing.
- M-PID: M-PID tries to adjust the bidding parameters according to the current bidding results. As shown by Fig. 2, compared with FP, M-PID is able to adjust parameters closer to the optimal ones therefore delivers better performance in dealing with environment change. However, there are two aspects of drawbacks that makes it less practical than RL methods: first, it is critical for M-PID to set a target for each indicator. However, setting an optimal target requires expert knowledge. As a result, the performance of M-PID is inferior than RL methods, i.e. USCB. Second, the parameter search process of this method is inefficient. We need to perform a grid search on all hyper-parameters. The size of the search space is exponentially related to the number of hyper-parameters. This will also decreases the practicality of M-PID when the number of constraints increases.
- DRLB: DRLB is also a model-free RL method and is proved to theoretically converge to the optimal policy. Therefore it delivers better performance than FP and M-PID, which shows the superiority of RL modeling. The rewardNet proposed in DRLB partially solves the problem of reward fluctuation. However, the reward is still affected by the episode that has already been learned, which makes it more noisy than USCB. As a result, it converge slower than USCB in practice. We will discuss about the convergence of these two methods in Section 4.3.
- USCB: Compared with the above methods, USCB is a RL method that is trained to learn the adjustment of bidding parameters at various state and is largely simplified according to Thm. 3.1. As illustrated in Fig. 2, USCB is able to quickly and steadily adjust bidding parameter close to the optimum. Therefore it delivers the best performance in all datasets and in all deviations. Since it is simple, efficient and effective, it can be widely applied to industrial constrained bidding problems.

4.3 Converging Efficiency

In industrial application, for technical iteration and cost savings, it is necessary for a method to accomplish the training process in an acceptable period of time and take up less computing resources during deployment. Methods based on reinforcement learning usually suffered from converging efficiency problems. In order to compare the converging efficiency of USCB and DRLB, the average metric G on all campaigns during training process is illustrated in Fig. 3. It can be seen that USCB converges faster than DRLB in all datasets. Specifically, on the dataset of CB{conversion-CPC}, USCB achieves a result of G higher than 0.9 using 1000 training episodes while DRLB requires about at least 2000 training episodes.

The outstanding converging efficiency of USCB should owe to the simplification of learning process based on Thm. 3.1. In actor-critic framework (the method DDPG [18] applied in our paper is based on actor-critic architecture), the accuracy of critic Q will directly affect the quality of the actor [20]. According to Thm. 3.1, at a specific state s_t, the best action is to adjust the current parameters

to the optimal ones and keeps them fixed to the end of an episode. Therefore, logically, the best action under s_t should be optimally criticized by critic Q. Recall that we let Q learn $\mathcal{G}$ in Algo. 1, and $\mathcal{G}$ is strongly related to the core metric we want to be optimized in application. The distributions of $\mathcal{G}$ over different parameters would directly affect the learning process of Q. In order to empirically show how Q is learned, we illustrate the distribution of $\mathcal{G}$ over different parameters (actions) at state s_1 of two specific episodes of CB{click} and CB{conversion-CPC} in Fig. 4. We can see that the better actions (adjusting parameters closer to optimal parameters) will receive a better $\mathcal{G}$. Therefore, the critic will quickly learn to differentiate which action is better to $\mathcal{G}$ and, in turn, boost the actor converging.

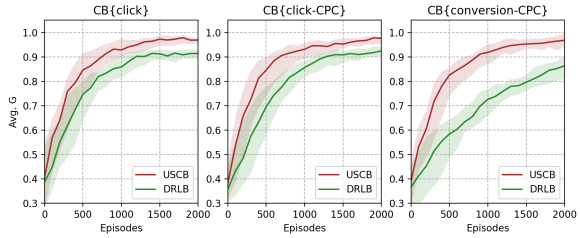

Figure 3: The training process of USCB and DRLB for the three constrained bidding problems.

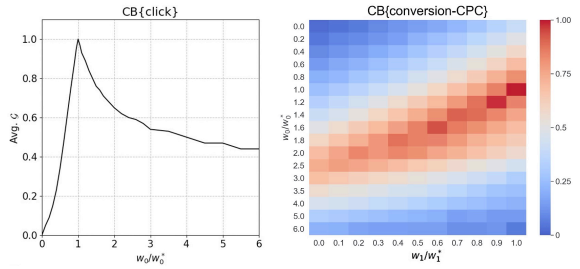

Figure 4: An illustration of the $\mathcal{G}$ distributions of critic Q for single parameter problem CB{click} and double parameters problem CB{conversion-CPC} at a starting state s_1. Suppose w_0 and w_1 are parameters having adjusted and the w_0^* and w_1^* are the optimal ones. We use a thermodynamic chart to express the tripartite relationship of w_0/w_0^*, w_1/w_1^* and $\mathcal{G}$. It can be seen that the closer of the parameters to the optimal ones, the better $\mathcal{G}$ we get.

4.4 Discussion on the Effect of Constraint Violation Penalty

In application, since the future impression set is hard to foresee, the parameter adjustment policy indeed could cause KPI constraint violation, therefore, we put a constraint violation penalty term in metric G. It is worth noting that G not only criticize the empirical result, but also affects the training process of a RL model. Specifically, the effect of the hyper-parameter λ in Eq. (17) to model training is worth investigating. We extract campaigns whose optimal KPIRatio is 1.0 (these campaigns are more prone to constraint violation and the behaviors of RL model would be more vivid), and then depict the metric G, KPIRatio and R/R^* over λ in Fig. 5 for CB{click-CPC} and CB{conversion-CPC}. In these cases, the KPIRatio is the actual CPC of all winning impressions (i.e. $\Sigma_i c_i x_i / \Sigma_i \text{pCTR}_i x_i$) divided by the CPC upper bound provided by the advertiser. It can be seen that USCB achieves a satisfying result (avg. G higher than 0.95) with almost all λ, i.e. USCB is robust to the hyper-parameter λ. What is more, when λ is small, the model will tend to be aggressive and has greater possibilities in KPI violation. As λ increases, the KPIRatio get smaller, which means the model will more care about constraint satisfaction. This shows that λ has a great influence on the agent's strategic tendency and can be an important tool for us to trade off between the constraints fulfillment and value acquisition.

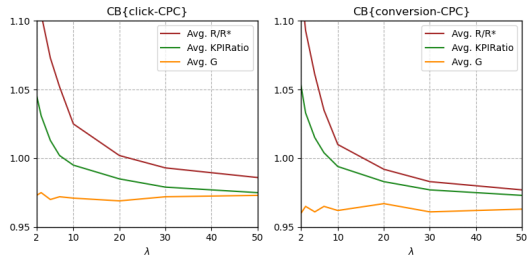

Figure 5: The illustration of RL model behaviors from the perspective of penalty strength for CB{click-CPC} and CB{conversion-CPC}. λ has a great influence on the agent's strategic behaviors and can be an important tool to trade off between the constraints fulfillment and value acquisition.

4.5 Online Deployment of USCB

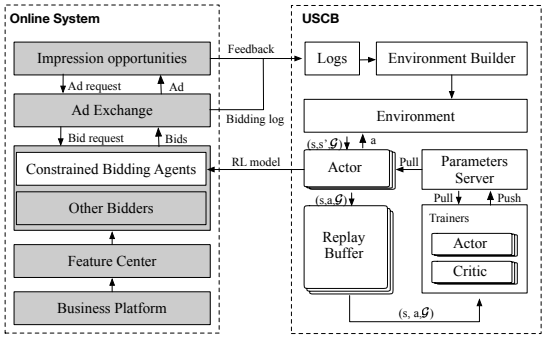

Figure 6: Illustration of USCB deployment in Taobao advertising platform.

The unified solution to constrained bidding (USCB) has been deployed and justified in Taobao display advertising system. Our method serves thousands of advertisers and impacts millions of revenue every day. The architecture of the algorithm application is shown in the Fig. 6. The training and deployment of the model are parallelized, which allows the model to be iterated efficiently and can be easily applied to a large number of advertising campaigns.

5 RELATED WORK

The bid optimization is a widely studied problem in RTB of online display advertising. [31] model BCB as an online knapsack problem. [19] proposed static bid optimization solutions. [28] set bids according to the static distribution of input data. In [13], MDP is used to perform online decision in tuning the impression level bid price. [5] proposed a reinforcement learning approach that shows robustness to the non-stationary auction environment. Some work has been proposed to specifically address the KPI constraints. [17] introduced a generic bidding solution to take into consideration of the advertising value and multiple KPI constraints. [25] utilized feedback control theory to address the dynamic environment. Different from previous works, our work was not limited to a few constraints or optimization goals. We abstracted the constrained bidding problem as a linear programming problem and derived the optimal bidding strategy through the primal-dual method which is generally applied in the ad allocation scenario [1, 6, 11]. To address the unstable traffic problem, we took advantage of the reinforcement learning, which has been proved effective in many scenarios by work of [15, 23, 29]. Moreover, impression value estimation play a very important role in RTB scenarios, e.g. click-through rate (CTR) [30] and conversion rate (CPI) [24], which helps to bid in the impression level precisely.

6 CONCLUSIONS

In this paper, we propose a unified solution to solve the constrained bidding problem in online display advertising. We first abstract the core demand of advertisers and formulate it as a constrained bidding problem. Then we leverage the primal-dual method to derive the optimal bidding function under second price auction. The optimal function allows advertisers calculate bids for all impressions with limited number of parameters. In order to address the challenge of non-stationary environment which results in deviation of parameters between consecutive days, we further propose a RL method to dynamically adjust parameters to the optimal ones. What is more, we found that the constrained bidding problem is a recursive optimal one, and this property significantly boosts the converging process of the learning process. Comprehensive experiments are conducted to verify the effectiveness of our solution. Our formulation and the RL method, together, is called Unified Solution to Constrained Bidding (USCB), which is deployed and justified in Taobao advertising platform.

REFERENCES

[1] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. 2014. A dynamic near-optimal algorithm for online linear programming. *Operations Research* 62, 4 (2014), 876–890.

[2] alibaba. 2021. Alimama Super Diamond. https://zuanshi.taobao.com/.

[3] Azarnoush Ansari, Arash Riasi, et al. 2016. An investigation of factors affecting brand advertising success and effectiveness. *International Business Research* 9, 4 (2016), 20–30.

[4] Achim Bachem and Walter Kern. 1992. Linear programming duality. In *Linear Programming Duality*. Springer, 89–111.

[5] Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. 2017. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 661–670.

[6] Ye Chen, Pavel Berkhin, Bo Anderson, and Nikhil R Devanur. 2011. Real-time bidding algorithms for performance-based display ad allocation. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1307–1315.

[7] Vasek Chvatal, Vaclav Chvatal, et al. 1983. *Linear programming*. Macmillan.

[8] A Ebrahimnejad and SH Nasseri. 2009. Using complementary slackness property to solve linear programming with fuzzy parameters. *Fuzzy Information and Engineering* 1, 3 (2009), 233–245.

[9] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review* 97, 1 (2007), 242–259.

[10] facebook. 2021. Advertising on Facebook. https://www.facebook.com/business/ads.

[11] Ashish Goel, Mohammad Mahdian, Hamid Nazerzadeh, and Amin Saberi. 2010. Advertisement allocation for generalized second-pricing schemes. *Operations Research Letters* 38, 6 (2010), 571–576.

[12] Google. 2021. Google Ads. https://ads.google.com/.

[13] R Gummadi, Peter B Key, and Alexandre Proutiere. 2011. Optimal bidding strategies in dynamic auctions with budget constraints. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 588–588.

[14] IAB. 2020. Full-Year 2019 Internet Advertising Revenue Report and Coronavirus Impact on Ad Pricing Report in Q1 2020. https://www.iab.com/video/full-year-2019-internet-advertising-revenue-report-and-coronavirus-impact-on-ad-pricing-report-in-q1-2020/.

[15] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. 2018. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2193–2201.

[16] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[17] Brendan Kitts, Michael Krishnan, Ishadutta Yadav, Yongbo Zeng, Garrett Badeau, Andrew Potter, Sergey Tolkachov, Ethan Thornburg, and Satyanarayana Reddy Janga. 2017. Ad Serving with Multiple KPIs. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1853–1861.

[18] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[19] Claudia Perlich, Brian Dalessandro, Rod Hook, Ori Stitelman, Troy Raeder, and Foster Provost. 2012. Bid optimizing and inventory scoring in targeted online advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 804–812.

[20] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[21] Jun Wang and Shuai Yuan. 2015. Real-time bidding: A new frontier of computational advertising research. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 415–416.

[22] Christopher A Wilkens, Ruggiero Cavallo, Rad Niazadeh, and Samuel Taggart. 2016. Mechanism design for value maximizers. *arXiv preprint arXiv:1607.04362* (2016).

[23] Di Wu, Xiujun Chen, Xun Yang, Hao Wang, Qing Tan, Xiaoxun Zhang, Jian Xu, and Kun Gai. 2018. Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1443–1451.

[24] Jia-Qi Yang, Xiang Li, Shuguang Han, Tao Zhuang, De-Chuan Zhan, Xiaoyi Zeng, and Bin Tong. 2020. Capturing Delayed Feedback in Conversion Rate Prediction via Elapsed-Time Sampling. *arXiv preprint arXiv:2012.03245* (2020).

[25] Xun Yang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. 2019. Bid optimization by multivariable control in display advertising. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1966–1974.

[26] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. 2013. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*. 1–8.

[27] Weinan Zhang, Kan Ren, and Jun Wang. 2016. Optimal real-time bidding frameworks discussion. *arXiv preprint arXiv:1602.01007* (2016).

[28] Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1077–1086.

[29] Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. 2018. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1021–1030.

[30] Guorui Zhou, Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Na Mou, Xinchen Luo, et al. 2020. CAN: Revisiting Feature Co-Action for Click-Through Rate Prediction. *arXiv preprint arXiv:2011.05625* (2020).

[31] Yunhong Zhou, Deeparnab Chakrabarty, and Rajan Lukose. 2008. Budget constrained bidding in keyword auctions and online knapsack problems. In *International Workshop on Internet and Network Economics*. Springer, 566–576.