An Adaptive Unified Allocation Framework for Guaranteed Display Advertising

Xiao Cheng[1], Chuanren Liu[2*], Liang Dai[3], Peng Zhang[1], Zhen Fang[3], Zhonglin Zu[3]
[1] Alibaba Group, Beijing, China, {jane.cx,freeman.zp}@alibaba-inc.com
[2] The University of Tennessee, Knoxville, USA, cliu89@utk.edu
[3] Alibaba Group, Hangzhou, China, {dailiang.dl,fangzhen.pt,zhonglin.zuzl}@alibaba-inc.com

ABSTRACT

Guaranteed Display (GD) is widely used in e-commerce marketing for advertisers to acquire an agreed-upon number of impressions with target audiences. With the main objective to maximize the contract delivery rate under contract constraints, user interest (such as click-through rate and conversion rate) is also essential to improve the long-time return on investment for advertisers and the e-commerce platform. In this paper, we design an adaptive unified allocation framework (AUAF) by not only considering supply of audience impressions in request-level but also avoiding over-allocation of audience impressions. Specifically, our allocation model simultaneously optimizes the contract delivery and the match between advertisements and user interests with explicit constraint to prevent unnecessary allocation. Facing the challenge of serving billion-scale requests per day, a parameter-server based parallel optimization algorithm is also developed, enabling the proposed allocation model to be efficiently optimized and incrementally updated in minutes. Thus, the offline optimization results and the online decisions can be synchronized for real-time serving. In other words, our approach can achieve adaptive pacing that is consistent with the optimal allocation solution. Our extensive experimental results demonstrate that the proposed AUAF framework can improve both contract delivery rate and average click-through rate (CTR), which we use to measure the user interest in this paper. The improvements on CTR are statistically significant in comparison with existing methods. Moreover, since March 2020, AUAF has been deployed in the guaranteed display advertising system of Alibaba, bringing more than 10% increase on CTR without loss of contract delivery rate, which has resulted in significant value creation for the business.

CCS CONCEPTS

• **Information systems** → **Computational advertising**; **Display advertising**.

KEYWORDS

E-commerce Marketing; Guaranteed Display Advertising; Optimal Allocation; Computational Advertising.

ACM Reference Format:
Xiao Cheng[1], Chuanren Liu[2*], Liang Dai[3], Peng Zhang[1], Zhen Fang[3], Zhonglin Zu[3] . 2022. An Adaptive Unified Allocation Framework for Guaranteed Display Advertising . In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3488560.3498500

1 INTRODUCTION

Online display advertising has become an important revenue source for e-commerce platforms and attracted much research attention [8, 20]. Guaranteed Display (GD) is a specific form of online display advertising in which advertisers and publishers sign contracts for a certain number of advertisement impressions in advance. Besides impressions, payment amount, campaign duration, target audience, frequency and reach requirements are also clarified in the contract. Moreover, penalty is charged to publishers if the impression demand is not fully satisfied.

For advertising publishers to achieve maximal revenue with GD advertising contracts, the major objective is to meet the impressions required by advertisers. Besides, publishers also try their best to improve the user interest of GD advertisements, where the user interest is measured by Click-Through Rate (CTR) and Conversion Rate (CVR) [8, 10, 13]. High user interest can improve attractiveness of the publisher compared to other competing publishers, and help to establish long-time cooperative business relationship with advertisers. Therefore, considering and optimizing user interest are important tasks in GD advertising.

To achieve these goals, a GD allocation model can be formulated as an resource allocation problem based on a bi-graph, where user crowds, individuals, or requests are supply nodes and advertisement contracts are demand nodes [2, 4, 10]. For instance, if one supply node in the bi-graph represents a user crowd, the edge between one demand node and one supply node can indicate that the user crowd is the target audience of the contract. Moreover, to cope with the time and data gaps between online serving and offline optimization, the GD allocation process can also consider an adaptive pacing strategy for online serving and real-time delivery [3, 18, 22].

However, GD optimization is non-trivial for large publishers if individual users or requests are considered as supply nodes, in which case, the scale of the bi-graph and the allocation model can reach billions. Specifically, there are two major challenges still not fully addressed in existing approaches:

* Corresponding author.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA

ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3498500

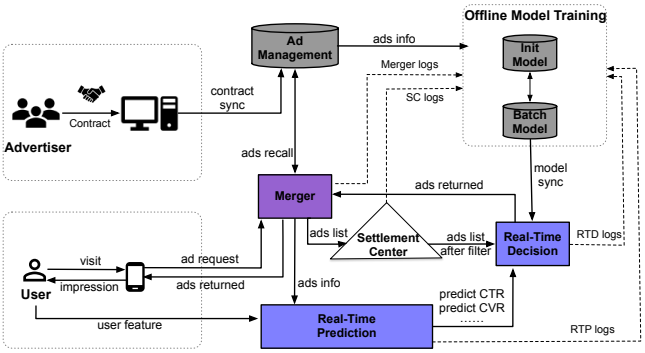

Figure 1: An overview of GD advertising system.

- To avoid penalty due to unsatisfied demand, existing GD allocation models usually result in solutions with over-allocation, which means the allocated impressions can be more than the required amount. Although such over-allocation will be eventually filtered out in practice, the allocation decisions based on the imperfect solutions may lead to difficulties in serving other contracts and maximizing user interests.
- Though pacing strategy in real-time serving and delivery plays an important role to adjust dynamic changes of supply and demand levels, this adjustment is likely to introduce biases to the allocation solution computed from the offline allocation model. Moreover, since online adjustment is usually independent for each contract, interactions among related (e.g., competing) contracts are not effectively considered to globally improve the overall allocation performance.

To address these challenges, in this paper, we consider a guaranteed display allocation system to avoid over-allocation in GD optimization and efficiently support online serving. As shown in Figure 1, GD contracts are signed with advertisers. Contract information are stored in the Ads Management Center (AMC), including ad index, required impressions, target audience, campaign duration, and frequency requirements. In online serving phase, related ads are recalled from AMC when a request arrives. Then those ads will be filtered by the Settlement Center (SC) if their current impressions have exceeded the demands in contracts. For each advertisement recalled in a request, user interest scores are estimated in the Real-Time Prediction (RTP) module. Finally, a certain number of ads are selected in the Real-Time Decision (RTD) module based on the allocation model, which choose ads to be displayed for the request. Meanwhile, all these above information are logged and transferred to the Offline Model Training (OMT) module. An optimal allocation solution is computed by the OMT module and updated to the RTD module in minutes to support online serving and delivery.

Based on the above GD allocation system, this paper focuses on the allocation model optimization in the OMT module and the online serving in the RTD module. Our focus is to simultaneously improve delivery rate and increase user interest by designing an adaptive unified allocation framework (AUAF). The framework consists of an over-allocation preventing model and an efficient dual-based parallel algorithm adaptive to dynamic online environments. The contributions of this paper can be summarized as follows:

- We propose a request-level user interest-oriented GD allocation model with the objectives of maximizing both delivery rate and user interest. We consider the demand capacity constraint in this model to prevent over-allocation.
- We propose a GD delivery framework that unifies offline optimization and online serving. For offline optimization, a dual-based algorithm is developed to obtain feasible allocation solutions. For online serving, near real-time decision is carried out according to the offline solution. Particularly, the offline optimization results and the online decisions can be synchronized for online serving. In other words, our approach can achieve adaptive pacing which is consistent with the optimal allocation solution.
- Since March 2020, the proposed adaptive unified allocation framework (AUAF) has been deployed in the guaranteed display advertising system of Alibaba. Extensive online A/B testing validates the effectiveness and efficiency of our framework. In comparison with latest solution before the implementation of our framework, AUAF brings more than 10% improvement on CTR under the same contract delivery rate, which has resulted in significant business value creation for the advertising system.

The rest of this paper is organized as follows: In section 2, related works on GD allocation and online real-time control are discussed. The basic model and our research questions are described in Section 3. In Section 4, we introduce our framework including the over-allocation-preventing model and the dual-based parallel algorithm. Experiments with both offline and online A/B testing are presented in Section 5. Finally we conclude our work in Section 6.

2 RELATED WORK

We review two categories of existing work: guaranteed display advertising allocation and real time feed-back control.

Guaranteed display advertising allocation. The GD allocation problem is often considered as weighted matching problem to maximize the weighted match between bipartite graph nodes [2, 12, 15, 16]. Assuming user traffic arrives in a random order, Feldman et al. [9], Karande et al. [11] showed that the greedy algorithm can achieve the approximation ratio of $1 - 1/e$, and Devanur and Hayes [7] introduced an offline-training phase with a primal-dual framework which can solve the GD allocation problem. Accordingly, some practical algorithms have been developed under the primal-dual framework. Vee et al. [17] utilized a particular subspace of the dual space and proposed a compact allocation plan to make near-optimal online decisions. Motivated by this idea, HWM [5] and SHALE [4] have been developed with improved efficiency and flexibility. Later, Hojjat et al. [10] proposed to use a column generation method to optimize the user interest under reach and frequency requirements. Zhang et al. [21] proposed a consumption minimization model where the primary objective is to minimize the user traffic consumed to satisfy all contracts. Lei et al. [13] used genetic algorithm with a specific design of coding scheme for video service platform. However, those approaches solve the GD allocation problem at crowd level of user traffic. To handle more fine-grained user interest and contract requirements, Fang et al. [8] and Zhang et al. [20] proposed distributed algorithms at user

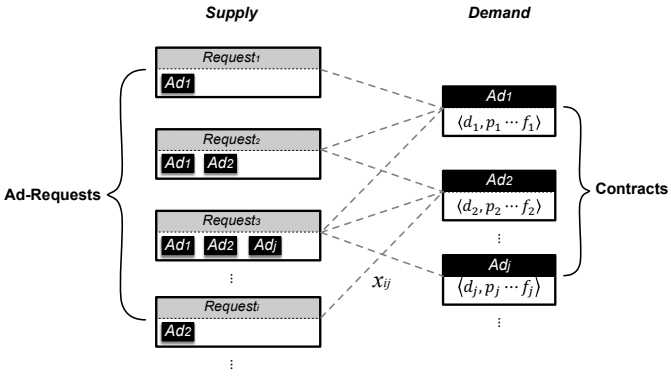

Figure 2: Bi-graph of GD allocation.

or request level. Nonetheless, there exists some gaps between the offline model assumption and the actual online allocation which limit their capacity of meeting more user interest.

Real time feed-back control. Advertisers prefer to spend their budget smoothly over the delivery period to reach a wide range of audience. It is natural to use a feedback controller to monitor the difference between actual and expected budget consumption and smoothly control the ad allocation. To this end, various dual-based pacing systems have been introduced to tune smoothing parameters. For instance, Chen et al. [6] used a linear programming (LP) primal-dual formulation to adjusts bids against real-time constraint satisfaction levels. Bhalgat et al. [3] introduced the online computation of dual variables to reach a wider range of audience. Under the same framework, Yang et al. [19] proposed a feedback control-based solution and designed their multi-variable control system. It has also been shown that offline data can be leveraged to improve pacing strategies. To leverage both offline and online data, Xu et al. [18] proposed to adjust the pacing rate with different optimization goals based on real-time feedback and offline flow distribution.

3 ALLOCATION MODEL FOR GUARANTEED DISPLAY ADVERTISING

GD advertising contracts involve three participants: advertisers, publishers, and users (i.e., advertising audience). A contract is created between an advertiser and the publisher to guarantee a specified number of advertisement impressions during a future period. In addition to the required impressions, contracts also specify requirements on targeted audiences, frequency, and reach of the allocated impressions. The publisher delivers advertisements to users until the number of impressions meets the demands in the contracts. The publisher will pay penalty cost if some demands are not satisfied after the contract period. These concepts are illustrated by the bi-graph shown in Figure 2.

In addition to the contract requirements, publishers also want to optimize the advertisement effects which are beneficial to both advertisers and users. For instance, the allocation objective can incorporate user interests and maximize the matching between users and advertisements. Without loss of generality, we use CTR as the measure of the allocation match. The intuition is that if the allocation match users with interesting and relevant advertisements, we would observe high CTR following the GD allocation.

3.1 Basic Allocation Model

For the publisher, to obtain the maximal revenue, the number of delivered impressions must reach the demands as close as possible. Besides, for long-term profits, CTR should also be considered. What's more, impressions allocation fairness is another weakly objective. Thus the basic guaranteed display advertising allocation model [4, 8, 20] is formulated as:

$$\arg\min_{x_{ij},u_j} \frac{1}{2} \sum_{j,i\in\Gamma(j)} s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j)^2 + \sum_j p_j u_j - \sum_j \lambda_j \sum_{i\in\Gamma(j)} s_i x_{ij} c_{ij} \tag{1}$$

s.t.

$$\sum_{i\in\Gamma(j)} s_i x_{ij} + u_j \geq d_j, \quad \forall j \tag{2}$$

$$\sum_{j\in\Gamma(i)} x_{ij} \leq 1, \quad \forall i \tag{3}$$

$$x_{ij} \geq 0, \quad \forall i,j \tag{4}$$

$$u_j \geq 0, \quad \forall j \tag{5}$$

where x_{ij} indicates the allocation probability from request (i.e., supply) i to contract (i.e., demand) j, and c_{ij} represents the user interest (e.g., the CTR predicted by a deep neural network [8]) between the request and the demand. We omit the interest prediction details and focus on the allocation model in this paper. s_i is the capacity of request i; d_j is the required impressions of demand j; λ_j is the importance of user interest for demand j; u_j is the unsatisfied amount of impressions of demand j; and p_j is the unit penalty specified in the j-th contract.

The objective function Equation 1 includes three addend, where the first addend maximizes allocation fairness; the second addend minimizes the gap between required demands and allocated amount; and the third one maximizes the user interest [4]. Specifically, the fairness is measured by the weighted ℓ_2 between x_{ij} and the fair allocation probability $\theta_j = \frac{d_j}{\sum_{i\in\Gamma(j)} s_i}$, where V_j in the objective function is the importance of allocation fairness for the j-th contract.

The qualify (e.g., based on user interest) of the match between the request i and the contract j is predicted by a score c_{ij}. The edge between supply node i and demand node j in the bi-graph means the ad j is recalled by request i. The set $\Gamma(j)$ means all the requests that recall ad j, and the set $\Gamma(i)$ is all the ads recalled by request i.

3.2 Our Research Questions

We address two problems, namely, over allocation and online inconsistency, with our AUAF framework.

Over allocation: The basic allocation model penalizes u_j, the gap between the contract demand and the delivered impressions in Equation 1. However, there is no mechanism in the basic allocation model to avoid over allocation, i.e., the delivered impressions exceed the contract demand. Since the online serving phase usually filters out such extra impressions, the online serving solution based on the basic allocation may not effectively optimize the objective function anymore. Each impression opportunity filtered out because of over allocation is a loss of supply resource, which might be re-allocated

to alternative contracts. One of our major contributions in this paper is to avoid over allocation in the offline optimization phase and show that the new optimization can result in superior online serving performances.

Online inconsistency: In order to compensate for the delay of offline model computation and achieve smooth delivery during online serving, traditionally a pacing method such as Proportional Integral Derivative (PID) is desired. An adjusted CTR threshold is obtained by PID to drop out the poor quality requests. However, PID calculation and adjustment are based on the error in previous period of time. In addition, PID is a single-point pacing of a single demand node, and does not consider the competitive relationships among GD demand contracts. Thus this pacing method is mostly independent of the offline allocation model, and the optimal solution obtained by the offline model may not be fully utilized.

4 ADAPTIVE UNIFIED ALLOCATION FRAMEWORK

The allocation of guaranteed display advertising relies on two phases: offline model optimization and online serving. Our approach, an Adaptive Unified Allocation Framework (AUAF), aims to support both phases for GD advertising. First, an over-allocation preventing model is established to address the questions in Section 3. Particularly, in the offline phase, an optimal allocation solution is obtained from a global perspective. Then, during the online serving phase, when a request arrives, our approach can improve the user interest in the allocation by reducing the gap between online and offline data distributions.

4.1 Over-Allocation-Preventing Model

Our approach aims to prevent the allocation solution from delivering unnecessary audience impressions (e.g., more than the amount specified in GD contracts) to impression demands. We will show that the saving of audience impressions can be re-targeted with demands that are either more difficult to satisfy or more relevant with respect to user interests. Specifically, for the j-th contract with demand d_j, we formulate the limit on demand capacity as an optimization constraint in Equation 7. Our new optimization problem is:

$$\begin{aligned}\arg\min_{x_{ij}} \quad & \frac{1}{2}\sum_{j,i\in\Gamma(j)} s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j)^2 - \sum_j w_j \sum_{i\in\Gamma(j)} s_i x_{ij} \\ & - \sum_j \lambda_j \sum_{i\in\Gamma(j)} s_i x_{ij} c_{ij}\end{aligned} \tag{6}$$

s.t.

$$\sum_{i\in\Gamma(j)} s_i x_{ij} \le d_j, \quad \forall j \tag{7}$$

$$\sum_{j\in\Gamma(i)} x_{ij} \le 1, \quad \forall i \tag{8}$$

$$x_{ij} \ge 0, \quad \forall i,j \tag{9}$$

where w_j represents the importance of delivery requirement for the j-th contract. The objective function in Equation 6 also includes three addend, where the first addend and the last addend are the same with those in the basic allocation model. The second addend in our model is formulated to maximize the amount of impressions for all contracts, with the first constraint in Equation 7 preventing unnecessary allocations over the contract amount. The last two sets of constraints in Equation 8 and Equation 9 are the same with those in the basic model to ensure probabilistic allocation solutions.

The key difference in our model compared with the basic allocation model is that the original variable u_j representing the unsatisfied amount of impressions is removed to explicitly prevent unnecessary allocations. Of note, if the user interest parts are removed from both models, the two models are equivalent [4, 8]. However, with user interest incorporated into the objective functions, their optimization results will result in different allocation solutions and business implications.

4.2 Dual-based Parallel Optimization

Since our model in Equation 6-9 is formulated with a convex objective function and all linear constrains, we can use the KKT conditions to find the optimal solutions for our model. Specifically, the Lagrangian function is:

$$\begin{aligned}L(\alpha,\beta,\gamma) =& \frac{1}{2}\sum_{j,i\in\Gamma(j)} s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j)^2 - \sum_j w_j \sum_{i\in\Gamma(j)} s_i x_{ij} \\ & - \sum_j \lambda_j \sum_{i\in\Gamma(j)} s_i x_{ij} c_{ij} + \sum_j \alpha_j \Big(\sum_{i\in\Gamma(j)} s_i x_{ij} - d_j\Big) \\ & + \sum_i \beta_i \Big(\sum_{j\in\Gamma(i)} s_i x_{ij} - s_i\Big) - \sum_{j,i\in\Gamma(j)} \gamma_{ij} x_{ij}\end{aligned} \tag{10}$$

It follows that the KKT conditions are:

$$s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j) - w_j s_i - \lambda_j s_i c_{ij} + \alpha_j s_i + \beta_i s_i - \gamma_{ij} = 0 \tag{11}$$

$$\alpha_j \Big(\sum_{i\in\Gamma(j)} s_i x_{ij} - d_j\Big) = 0 \tag{12}$$

$$\beta_i \Big(\sum_{j\in\Gamma(i)} x_{ij} - 1\Big) = 0 \tag{13}$$

$$\gamma_{ij} x_{ij} = 0 \tag{14}$$

$$\alpha_j \ge 0, \beta_i \ge 0, \gamma_{ij} \ge 0 \tag{15}$$

where α_j, β_i, and γ_{ij} are the Lagrangian multipliers of constraints Equation 7, 8, and 9, respectively. According to the KKT conditions, the optimal allocation probability x_{ij} can be derived as:

$$x_{ij} = \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j - \beta_i}{V_j})\} \tag{16}$$

According to Equation 16, the decision variable x_{ij} depends on the dual vectors α_j and β_i. To obtain the optimal x_{ij}, the values of α_j and β_i need to be calculated in advance. Since the dimension of α and β equals to the number of demand nodes m and the number of supply nodes n, respectively, the total number of parameters is greatly reduced from the nm in original space to $n+m$ in dual space.

The dual vectors α and β can be solved using coordinate descent method. For example, to solve α_j, Equation 12 should be solved for every demand node j, and Equation 13 needs to be solved to obtain β_i for each supply node i. In practice, the number of demand nodes for every supply node i is usually less than ten hundred, thus the accurate solution of β_i can be quickly solved. However, for solving Equation 12, since there is usually billions of supply

nodes connected to a demand j, we adopt an efficient approximate iteration to obtain α_j, with the update formula as:

$$\alpha_j^{t+1} = \alpha_j^t - V_j(1 - \frac{d_j(\alpha^t)}{d_j}) \tag{17}$$

where t means the t-th iteration and $d_j(\alpha^t)$ represents the allocated impressions for demand j in t-th iteration.

Since α can be obtained according to Equation 17, we further adopt a Parameter-Server (PS) architecture [14] to effectively solve this allocation model in parallel. The pseudo-code of our parallel allocation algorithm is shown in Algorithm 1.

Algorithm 1 Parallel Optimal Allocation Algorithm.

```
function UpdateBeta                                   ▷ Worker
    Pull all α_j from server
    for i ← 0 to n do
        Update β_i with Equation: Σ_{j∈Γ(i)} x_ij = 1
        for j ← 0 to m do
            Update x_ij with new β_i
        end for
    end for
    Push all s_i x_ij to server
end function
function UpdateAlpha                                  ▷ Server
    Gether all s_i x_ij from worker
    for j ← 0 to m do
        Update α_j^{t+1} with Equation:
        α_j^{t+1} = α_j^t − V_j(1 − d_j(α^t)/d_j)
    end for
    Update all α_j to worker
end function
```

The following theorem shows that Equation 17 and Algorithm 1 will produce a feasible solution for our allocation model.

Theorem 1. *If α_j is initialized with $\alpha_j^0 = w_j + \lambda_j \max_{i\in\Gamma(j)}\{c_{ij}\}$, the intermediate variables x_{ij} of Algorithm 1 are all feasible solution for the over-allocation-preventing model.*

Let α_j^t and β_i^t denotes the value of α_j and β_i calculated in t-th iteration respectively. And define $d_j(\alpha^t)$ as:

$$d_j(\alpha^t) = \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^t - \beta_i^t}{V_j})\} \tag{18}$$

Before prove Theorem 1, we will first prove the following lemma:

Lemma 1. *For any iteration t, $d_j(\alpha^t) \le d_j$.*

Proof of Lemma 1. *A mathematical Induction method is applied to proof Lemma 1:*

(1) When $t = 0$, because of $\alpha_j^0 = w_j + \lambda_j \max_{i\in\Gamma(j)}\{c_{ij}\}$ and $\beta_i \ge 0$, it can be derived that:

$$\begin{aligned} d_j(\alpha^0) &= \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^0 - \beta_i^0}{V_j})\} \\ &\le \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j\} \le d_j \end{aligned} \tag{19}$$

(2) Assume $d_j(\alpha^t) \le d_j$ for t, then we will show that $d_j(\alpha^{t+1}) \le d_j$ for $t+1$. According to $d_j(\alpha^t) \le d_j$, we have

$$\begin{aligned} & d_j - d_j(\alpha^t) \\ &= \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^{t+1} - \beta_i^t}{V_j})\} \\ &- \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^t - \beta_i^t}{V_j})\} \\ &\le \sum_{i\in\Gamma(j)} s_i \theta_j(1 + \frac{\alpha_j^t - \alpha_j^{t+1}}{V_j}) \end{aligned} \tag{20}$$

It follows that:

$$\alpha_j^{t+1} \le \alpha_j^t - V_j(1 - \frac{d_j(\alpha^t)}{d_j}) \tag{21}$$

From Equation 21, α_j is monotone decreasing. Further in order to maintain the equality in Equation 13, β_i is monotone increasing, which implies $\beta_i^{t+1} \ge \beta_i^t$. For $t+1$:

$$\begin{aligned} d_j(\alpha^{t+1}) &= \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^{t+1} - \beta_i^{t+1}}{V_j})\} \\ &\le \sum_{i\in\Gamma(j)} s_i \max\{0, \theta_j(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j^{t+1} - \beta_i^t}{V_j})\} \le d_j \end{aligned} \tag{22}$$

□

Based on Lemma 1, we can prove Theorem 1.

Proof of Theorem 1. *The feasible solution means that all constraints in the allocation model are satisfied. For the supply capacity constraints, since β_i is calculated by a linear analytical method, this constraint will be satisfied. Also the non-negativity of x_{ij} is apparently established according to Equation 16. So we will only prove that the demand constraint Equation 7 is valid for every iteration in Algorithm 1. This validity is based on that facts that $d_j(\alpha^t)$ represents the allocated impressions for demand j after iteration t.* □

Based on Theorem 1, the model constraints will be satisfied in the process of Algorithm 1. In other words, a feasible allocation probabilities x_{ij} that meet the constraints can be obtained after any iteration. In practice, we adjust the number of iterations according to the computation time requirements.

4.3 Adaptive Online Pacing

To support online allocation, Lagrangian multipliers α_j for each demand j in the offline allocation solution will be updated to the online servers. When a user request i arrives, β_i and x_{ij} can be calculated according to Equation 13 and Equation 16 to support the allocation decisions. The adaptive pacing strategy is important for online allocation decisions to retain contract demands for future user requests with potentially higher qualities (as measured by c_{ij}). Generally such adaptive pacing strategy should result in smooth deliveries during each decision window. For instance, the widely used PID controller can quickly adjust online traffics. However, the adjustments in PID only consider the delivery level of each demand

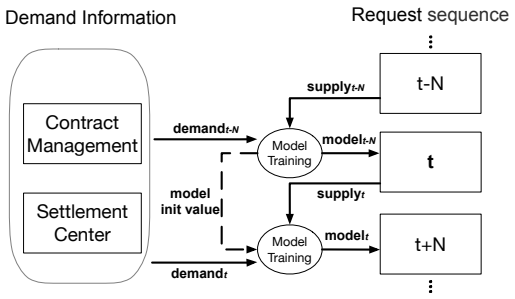

Figure 3: Diagram of GD allocation.

independently. In contrast, we aim at pacing requests from a global perspective that is consistent with the allocation model.

In this paper, our adaptive online pacing strategy is fully based on the allocation solutions (as denoted by x_{ij}). As mentioned in the previous section, the integer variable x_{ij} is approximated by stochastic variables. In other words, the x_{ij} obtained by solving the allocation model can be considered as an approximation of the 0-1 indicator, which can be interpreted as the probability that demand j should be displayed to the request i. Therefore, we can naturally filter out demand nodes with small allocation probabilities in online serving.

- Initialization: For the arriving request i, get every recalled demand j and corresponding α_j.
- Step 1: Calculate β_i using an efficient method according to Equation 13. Set $\beta_i = 0$ if there is no solution.
- Step 2: Calculate x_{ij} for every recalled demand j according to Equation 16.
- Step 3: Drop out the recalled demand j if $x_{ij} = 0$.
- Step 4: Select a specific amount of ads j to display for request i according to the stochastic value of x_{ij}, where the amount must be no more than the capacity of this request i.

The diagram of our proposed GD allocation is shown in Figure 3. Obviously, we have to update the offline allocation model frequently so that such stochastic pacing is effective and consistent with the offline solution. To this end, we use the incremental update method to speed up offline model optimization so that the allocation solution is synchronized to the online RTD module in minutes. Since the update speed of the offline model is very fast and the update frequency is high, the dynamic changes of the online serving environment can be captured and incorporated efficiently and effectively. Our experimental results show that our strategy can achieve the effect of adaptive pacing to the dynamic changes of the online serving environment. Meanwhile, by maintaining strong consistency with offline allocation solution, our strategy can achieve significantly better performance in balancing between delivery rate and user interest.

5 EXPERIMENTAL RESULTS

To evaluate the effectiveness of our approach (AUAF), we conduct several experiments on a large e-commerce marketplace. Several widely used methods are compared with ours. In the following, we first describe the data sets and introduce the benchmark methods. Then we discuss results in both offline and online experiments. Our code for AUAF and data sets are publicly available online[1].

5.1 Experiment Settings

5.1.1 Data sets and evaluation metrics. We use the data set from a sample of the user requests and advertisements in Alibaba's GD advertising platform, with more than 3 million supply nodes, 558 demand nodes, and more than 10 million edges in the allocation bi-graph per day. We also use a sample for simulations, with 125 thousand supply nodes, 128 demand nodes, and 298 thousand edges. As a benchmark, we use a linear programming (LP) optimizer [1] to obtain an upper bound of the GD allocation model, so the performance of different methods can be compared directly.

For the evaluation of guaranteed display advertising allocation, we consider the following two metrics:

- **Delivery rate**, also called contract completion rate, is the ratio of total valid allocation impressions to the required demand total. The valid allocation impressions don't include over-allocations so the delivery rate is at most 100%.
- **Click-through rate (CTR)** is the ratio of clicks to the allocated impressions. Since over-allocation will be filtered out in online serving, it doesn't contribute to click-through rate.

5.1.2 Benchmark methods. For the offline allocation model, we investigate the following three recent methods and compared them with our approach in experiments:

- SHALE: proposed by Bharadwaj et al. [4], is also a dual-based optimal algorithm for the basic allocation model which only considers two objectives of distribution fairness and maximum impressions.
- ALI: proposed by Fang et al. [8], is a an allocation model considering the optimization of CTR. However, there is no constraint restricting the allocated impressions cannot exceed the demands in the model. Instead, a hyper-parameter called learning rate is introduced in the update of α to retard over-loading during the iterations.
- RAP: proposed by Zhang et al. [20], is also based on an allocation model without restriction on over-allocation. We compute both the Lagrangian multipliers α and β by a gradient descent method.

For the online serving and pacing strategies, on the basis of above three offline optimization methods, we implement the following methods in addition to our approach for comparison: AUAF+PID, ALI+PID, and RAP+PID. To show the performance of incremental updating of our framework, We also compare the original AUAF and the incremental AUAF, IAUAF, that uses a faster incremental updating implementation. Important parameter values used in our experiments are tuned with grid search and summarized in Table 1.

[1] https://github.com/cxmlxq/AUAF.git

Table 1: Parameter Settings.

Parameter	Values	Description
w_j	100	Weight of impression objective
λ_j	100	Weight of user interest objective
v_j	1	Weight of fairness objective
p_j	100	Weight of penalty objective
lr	0.7	Learning rate in ALI
$t_{\max}$	100	Max iterations for all methods

5.2 Offline Allocation Evaluation

5.2.1 Model verification and correctness. We first conducted experiments to verify the correctness of our model and the proposed AUAF algorithm on the small-scale data set. The results are shown in Table 2, in which the upper bound is the optimal solution obtained by a direct linear programming optimization.

It can be demonstrated from Table 2 that the allocated impressions of different methods are all very close to the optimal value. For delivery rate, SHALE, ALI and AUAF show little difference compared with the upper bound. However, the delivery rate of RAP is slightly lower than that of others. For the over allocation rate, SHALE (without user interest in its objective function) and our AUAF result in no over-allocation, while ALI and RAP both have over allocation rate of more than 10%. This comparison reflects the advantages of our over-allocation preventing model. For the number of clicks and CTR, results of AUAF are closer to the upper bound than that of other methods. Overall, AUAF can achieve the CTR closest to the upper bound while maximizing delivery rate.

5.2.2 Large scale allocation evaluation. With the large-scale sample data, the results and convergence of the offline allocation model are shown in Figure 4. The final experiment results after the same number of iterations are summarized in Table 3. Our observations from the small-scale simulations are still valid in Table 3. The consistency can further confirm the reliability of our models. Moreover, we could run paired Student's t-test on the results from the large-scale data, by testing the difference between each method against AUAF and treating each GD contract as one sample in the t-test. According to the testing results, although delivery rate of SHALE is higher than that of AUAF, the difference is not significant with a p-value of 0.181. In contrast, the delivery rate of RAP is significantly lower than that of other approaches since the testing p-value is close to zero. In addition, due to the over-allocation that the basic allocation model cannot avoid, both ALI and RAP have more than 27% meaningless over-allocation. For click and CTR results, AUAF showed obvious advantages: compared with SHALE method without CTR objective, AUAF achieves 53% CTR increase; compared with ALI and RAP, AUAF increases the CTR by 19%. Such increases are all statistically significant: the p-values from paired t-tests between each method against AUAF are near zero.

Figure 4a, 4b, and 4c show the convergence of different methods during offline optimization. For the delivery rate, AUAF, ALI, SHALE can all converge in about 30 rounds. Among them, the initial rate of AUAF is low, but it can be quickly increased. The final result is consistent with Table 3: AUAF can achieve the most clicks. The convergence of CTR and delivery rate is inversely proportional, because these are two contradictory objectives. For the final CTR AUAF is still significantly better than other methods.

Figure 5a, 5b, and 5c are detailed data of the evaluation results in Table 3, including valid allocated impressions and CTR of different methods in the three sub-dimensions: advertising position, age, and province, respectively. The sub-dimensions in each figure are arranged from left to right according to the CTR in descending order. All the three figures can show two identical conclusions: 1) From the CTR in line, AUAF can achieve the maximum CTR in every sub-dimension; 2) From the valid allocated impressions in bar, the closer to the left dimension with high CTR, the more obvious impressions of AUAF is higher than other methods. This shows that AUAF tends to allocate advertisements to requests with higher CTR to achieve the goal of CTR maximization.

5.3 Online A/B testing

Finally, we evaluate the effectiveness of our approach by conducting online A/B testing for one week. Indeed, our method has been used in practice since March 2020, and our A/B testing aims to compare all related methods in a controlled and stable environment. Since the delivery rates of different methods are all above 99%, we will focus on the comparison of CTRs.

Figure 6a shows the predict CTR distribution of requests in different A/B testing groups. The little difference in the predict CTR distribution for different testing groups indicates the consistency of different methods in the A/B testing. Under this premise, we can compare the real CTR finally achieved by different methods.

Figure 6b shows the CTR achieved by different allocation strategies in online testing. The CTRs of the ALI and the RAP methods with online PID are basically the same, and the CTRs for both AUAF with and without PID are better than the previous two methods. This is due to the advantage of the optimization performance of AUAF. The comparison also shows that AUAF itself can achieve pacing without additional PID control. Moreover, IAUAF using incremental updating can achieve even higher CTR than AUAF because IAUAF can be more adaptive to online changes. In general, IAUAF can improve more than 10% on CTR without loss of delivery rate comparing with other methods in online A/B testing.

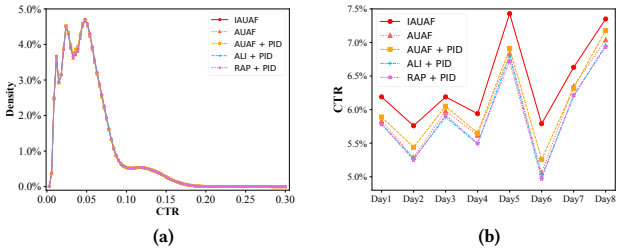

Figure 6: Predicted and Online CTR in A/B testing.

6 CONCLUSION

We present an adaptive unified allocation framework (AUAF) for guaranteed display advertising. First we propose an over-allocation-preventing model to maximizing delivery rate and user interest

Table 2: Simulation Result for Small-Scale Data Set.

Methods	Delivery impressions	Delivery rate	Over-allocation rate	Click	CTR
Upper Bound	**90313.00**	**0.639**	**0**	**4178.11**	**0.04626**
SHALE	90176.32	0.638	0	3929.59	0.04358
RAP	86931.44	0.615	0.1381	3800.90	0.04372
ALI	90213.18	0.638	0.1159	3952.62	0.04381
AUAF	90116.03	0.638	0	**4025.76**	**0.04467**

Table 3: Evaluation Result for Large-Scale Data Set.

Methods	Delivery impressions	Delivery rate / P-Value	Over-allocation rate	Click	CTR / P-Value
SHALE	1.207×10^6	**0.8237** / 0.181	0	3.27×10^4	0.0271 / 10^{-77}
RAP	1.198×10^6	0.8175 / 10^{-49}	0.2797	4.18×10^4	0.0349 / 10^{-71}
ALI	1.207×10^6	0.8236 / 0.255	0.2759	4.20×10^4	0.0348 / 10^{-75}
AUAF	1.207×10^6	**0.8237** / -	0	5.01×10^4	**0.0415** / -

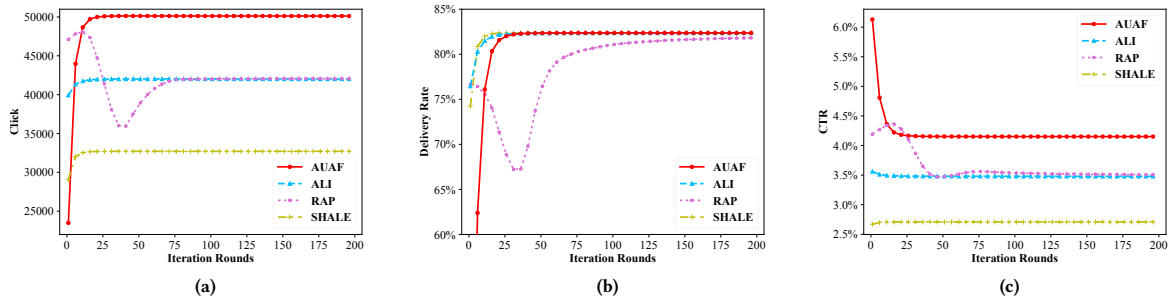

Figure 4: Results of large scale offline allocation.

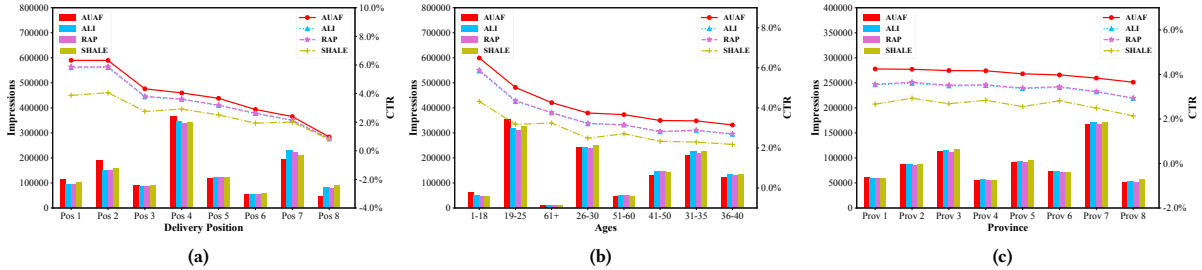

Figure 5: Allocation result in different contexts.

simultaneously, with consideration of both supply and demand resource constraints. We develop efficient dual-based parallel algorithm to solve the optimal allocation model, which can support online serving and adaptive pacing of the advertising system. Extensive experiments and online A/B testing show the advantages of our framework in improving both delivery rate and user interest. AUAF has been deployed in Alibaba's GD advertising system for more than one year and has brought considerable revenue improvement.

ACKNOWLEDGEMENTS

This work was supported by Alibaba Group through the Alibaba Innovative Research Program.

REFERENCES

[1] Google or-tools. URL https://developers.google.com/optimization/lp/glop.

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

[3] Anand Bhalgat, Jon Feldman, and Vahab Mirrokni. Online allocation of display ads with smooth delivery. In *KDD '12*, page 1213–1221, 2012. ISBN 9781450314626.

[4] Vijay Bharadwaj, Peiji Chen, Wenjing Ma, Chandrashekhar Nagarajan, John Tomlin, Sergei Vassilvitskii, Erik Vee, and Jian Yang. Shale: An efficient algorithm for allocation of guaranteed display advertising. In *KDD'12*, 2012. ISBN 9781450314626.

[5] Peiji Chen, Wenjing Ma, Srinath Mandalapu, Chandrashekhar Nagarjan, Jayavel Shanmugasundaram, Sergei Vassilvitskii, Erik Vee, Manfai Yu, and Jason Zien. Ad serving using a compact allocation plan. In *EC '12*, EC '12, page 319–336, 2012. ISBN 9781450314152.

[6] Ye Chen, Pavel Berkhin, Bo Anderson, and Nikhil R. Devanur. Real-time bidding algorithms for performance-based display ad allocation. In *KDD '11*, page 1307–1315, 2011. ISBN 9781450308137.

[7] Nikhil R. Devanur and Thomas P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *EC '09*, page 71–78, 2009. ISBN 9781605584584.

[8] Z. Fang, Y. Li, C. Liu, W. Zhu, Y. Zhang, and W. Zhou. Large-scale personalized delivery for guaranteed display advertising with real-time pacing. In *ICDM'19*, pages 190–199, Nov 2019.

[9] Jon Feldman, Aranyak Mehta, Vahab Mirrokni, and S Muthukrishnan. Online stochastic matching: Beating $1 - 1/e$. In *FOCS'09*, pages 117–126, 2009.

[10] S Ali Hojjat, John Turner, Suleyman Cetintas, and Jian Yang. Delivering guaranteed display ads under reach and frequency requirements. In *AAAI'14*, pages 2278–2284, 2014.

[11] Chinmay Karande, Aranyak Mehta, and Pushkar Tripathi. Online bipartite matching with unknown distributions. In *STOC '11*, page 587–596, 2011. ISBN 9781450306911.

[12] R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *STOC '90*, page 352–358, 1990. ISBN 0897913612.

[13] Hang Lei, Yin Zhao, and Longjun Cai. Multi-objective optimization for guaranteed delivery in video service platform. In *KDD '20*, page 3017–3025, 2020. ISBN 9781450379984.

[14] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 583–598, 2014.

[15] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5):22–es, October 2007. ISSN 0004-5411.

[16] Vahab S. Mirrokni, Shayan Oveis Gharan, and Morteza Zadimoghaddam. Simultaneous approximations for adversarial and stochastic online budgeted allocation. In *SODA '12*, page 1690–1701, 2012.

[17] Erik Vee, Sergei Vassilvitskii, and Jayavel Shanmugasundaram. Optimal online assignment with forecasts. In *EC '10*, page 109–118, 2010. ISBN 9781605588223.

[18] Jian Xu, Kuang-chih Lee, Wentong Li, Hang Qi, and Quan Lu. Smart pacing for effective online ad campaign optimization. In *KDD '15*, page 2217–2226, 2015. ISBN 9781450336642.

[19] Xun Yang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. Bid optimization by multivariable control in display advertising. In *KDD '19*, page 1966–1974, 2019. ISBN 9781450362016.

[20] Hong Zhang, Lan Zhang, Lan Xu, Xiaoyang Ma, Zhengtao Wu, Cong Tang, Wei Xu, and Yiguo Yang. A request-level guaranteed delivery advertising planning: Forecasting and allocation. In *KDD '20*, page 2980–2988, 2020. ISBN 9781450379984.

[21] Jia Zhang, Zheng Wang, Qian Li, Jialin Zhang, Yanyan Lan, Qiang Li, and Xiaoming Sun. Efficient delivery policy to minimize user traffic consumption in guaranteed advertising. In *AAAI'17*, pages 252–258, 2017.

[22] Weinan Zhang, Yifei Rong, Jun Wang, Tianchi Zhu, and Xiaofan Wang. Feedback control of real-time display advertising. In *WSDM '16*, page 407–416, 2016. ISBN 9781450337168.