# Integrated Ad Delivery Planning for Targeted Display Advertising

Huaxiao Shen
Business School, Sun Yat-sen University, Guangzhou 510275, China
hxshen@outlook.com

Yanzhi Li, Youhua (Frank) Chen
College of Business, City University of Hong Kong, Hong Kong
{yanzhili, youhchen}@cityu.edu.hk

Kai Pan
Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Hong Kong
kai.pan@polyu.edu.hk

Consider a publisher of online display advertising that sells its ad resources in both an upfront market and a spot market. When planning its ad delivery, the publisher needs to make a trade-off between earning a greater short-term profit from the spot market and improving advertising effectiveness in the upfront market. To address this challenge, we propose an integrated planning model that is robust to the uncertainties associated with the supply of advertising resources. Specifically, we model the problem as a distributionally robust chance-constrained program. We first approximate the program by using a robust optimization model, which is then transformed into a linear program. We provide a theoretical bound on the performance loss due to this transformation. A clustering algorithm is proposed to solve large-scale cases in practice. We implement ad serving of our planning model on two real data sets, and we demonstrate how to incorporate realistic constraints such as exclusivity and frequency caps. Our numerical experiments demonstrate that our approach is very effective: it generates more revenue while fulfilling the guaranteed contracts and ensuring advertising effectiveness.

*Key words*: display advertising, ad delivery planning, distributionally robust chance-constrained optimization, targeted advertising

## 1. Introduction

Online display ads can range from simple text, image, flash, and video to rich media, and such ads generally appear together with the contents of web pages. Compared with traditional advertising media such as radio and TV, online advertising offers a better audience targeting capability. Consequently, online display advertising has become a multi-billion dollar business. For example, according to the data of eMarketer (2019), the worldwide online display advertising revenue reached $144.7 billion in 2018, representing a 26.8% increase from 2017, and it will exceed $200 billion in 2020.

### 1.1. The Current Practice

The present study is motivated by the current industry practice. Specifically, we observe the ad delivery practices of two companies: a leading social network service (SNS) firm and a video advertising operator. Despite some expected differences in their operations, these companies face some common challenges. Below, we first detail the typical industry practice by referring to the SNS firm.

User visits to the publisher's website generate advertising opportunities that are subject to uncertainties. These uncertainties are derived from the variations in the number of users visiting the website, the number of times that the same user visits the website in a day, and the duration of each visit. Furthermore, users, or user visits, are characterized by various features, e.g., gender, age, education, occupation, and region, that allow advertisers to precisely target their preferred audiences. Such features information can be collected by the publisher itself and may also be purchased from third-party data management platform (DMP), such as Oracle and Adobe (Market Research Future 2019).

The publisher sells its advertising resources via either an upfront market or a spot market. In the upfront market, the publisher enters into a contract with each advertiser, guaranteeing an agreed-upon number of impressions over a specified period at a fixed price. These contracts often vary in the length of the period set for the ads; therefore, at any given time, the publisher has many contracts, each of which has a varying time span. In the spot market, the publisher auctions off its remaining advertising resources in real time among advertisers who, either purposefully or spontaneously, want to display their ads during the current period.

Therefore, at a higher level, over time, the publisher needs to decide whether to enter into guaranteed contracts with its advertisers and, if so, on what terms. Such decisions need to consider issues such as future advertising resource availability and possible revenue from the spot market. At the operational level, with the guaranteed contracts in place, the publisher needs to plan its ad delivery for each period. As its current practice, the publisher manages the two markets separately, with the upfront market being managed by the sales department and the spot market being managed by the (technical) operations department. Regarding ad delivery, to ensure that the guaranteed ads are fulfilled, the upfront market is given a higher priority, and ad resources are pre-allocated to the guaranteed ads. The remaining resources are then sold at the spot market via auction.

Although the above description considers the practice of the SNS firm, this description is rather typical and basically also applies to the video advertising firm. Large publishers attract a considerable amount of web traffic and thus have incentives to sell their advertising resources directly, whereas many other websites receive less traffic and may sell their advertising resources via ad exchange platforms, such as Google's DoubleClick (Balseiro et al. 2014). In this case, the problem studied in this paper is more suitable for an ad exchange platform.

### 1.2. Inefficiencies of the Current Practice

The current practice described above suffers from two major inefficiencies. The first is that separate management of the upfront market and spot market ignores the value of coordination and integration and thus misses potential profit. When a user visits a website, due to his/her multidimensional characteristics, a number of ads, from either the upfront or spot market, may set him/her as a target. In the current practice, an advertiser from the spot market may not receive the opportunity to display its ads even if it is willing to pay more because guaranteed ads are given higher priorities in ad delivery. Different ad campaigns may value the same user differently, and deliberately designed ad delivery can better exploit the value of advertising resources.

The second inefficiency is related to the selling approach in the spot market. Although theoretically appealing, simple auction mechanisms (e.g., the generalized second-price auction) have been found to have many drawbacks in practice. Realistic concerns such as daily budgets set by advertisers, advertisers' preference over smooth delivery, and different risk preferences among advertisers have made the theoretical appeal of the mechanisms doubtful. In fact, Benisch et al. (2009) proved that simple auction mechanisms can be arbitrarily inefficient in the presence of advertisers' budget constraints. Lu et al. (2015), through a game-theoretic analysis, found that the publisher's revenue can decrease due to advertisers' budget constraints. Similar to the results reported by earlier authors (Zhang and Feng 2011), we have found that advertisers do not change their bids frequently, although they do adjust over time, often daily. In other words, the publisher typically holds perfect or sufficient information about advertisers' bids that allows it to *plan* ad delivery over a period. Indeed, some authors from academia, such as Walsh et al. (2010), and from industry, such as Yang et al. (2010), have proposed adopting a planning approach rather than auctions to allocate the ads among advertisers.

### 1.3. Our Method and Contributions

In this paper, we propose integrating ad delivery planning for guaranteed ads and spot market ads. Specifically, we study the operational decisions of an online advertising publisher, namely, how the publisher should allocate its advertising resources to different guaranteed ads and to spot market ads, considering the uncertain supply of ad resources, the requirement of guaranteed ads, and the bids submitted from advertisers in the spot market.

As mentioned previously, each upfront contract covers a period ranging from days and weeks to months. Advertisers typically prefer smooth delivery of the ad campaign, meaning that the contracted number of impressions rolls out smoothly over time (Bhalgat et al. 2012). This allows the firm to subdivide the contract into shorter periods (e.g., a day) and plan the ad delivery for such refined periods. Thus, in each (refined) period, the firm needs to plan its ad delivery to fulfill its guaranteed contracts and subsequently maximize revenue from the spot market.

A number of concerns arise from the publisher's planning process. First, it is important to advertisers that the contracts be fulfilled for an ad campaign to achieve its objective. For example, a smart phone manufacturer launching a new product in October may purchase 100 million impressions from users aged 18 to 60 years in the U.S. from August to September. Achieving the impression target is critical to brand awareness and to the success of the new product launch, and the publisher is obliged to deliver the contracted number of impressions.

Second, rather than having its ads displayed to the same group of users repeatedly, an advertiser prefers to display the ads to as many unique targeted users as possible. The measure *reach* is defined as the number of unique individuals that an ad campaign has covered. Note that a user may visit the website several times during a period and that visits vary in duration. Thus, one user may generate multiple impressions during a period. A higher reach is critical to advertising effectiveness and is thus highly regarded by advertisers. Therefore, even if the guaranteed contract between the publisher and advertiser does not specify concrete terms regarding reach, to maintain a long-term relationship with the advertisers and build up its reputation, the publisher aims to improve along such a dimension.

Third, given all of the above, considering the limited advertising resources and the fixed revenue from guaranteed ads, the publisher has a natural motive to allocate more resources to the spot market to achieve a higher revenue. Therefore, the firm essentially faces a trade-off between earning a larger revenue in the spot market and improving the effectiveness of guaranteed ads to maintain a long-term relationship with advertisers.

In this paper, we propose an integrated planning model for ad delivery of both guaranteed and non-guaranteed ads. The proposed model incorporates all of the earlier concerns and allows the publisher to balance the trade-off between short-term profit and long-term benefit. Specifically, we make the following contributions.

(1) We derive a closed-form expression for the expected reach and then construct a distributionally robust chance-constrained (DRCC) programming model that maximizes the spot-market revenue subject to the attainment of contracted delivery and higher reach. The solution to our model is an ad delivery plan that is robust to the uncertainties of advertising resources.

(2) To solve the DRCC program, we propose a robust optimization model to safely approximate the original model, and the robust optimization model is further transformed into a linear program for tractability. We provide theoretical guarantees for the performance loss due to the transformation.

(3) We propose a clustering algorithm to efficiently solve large-scale instances of the linear program. In practice, the problem is often of a large scale and is challenging to solve. For example, if the audience is categorized by traits including 50 regions, 2 genders, 10 age groups, 5 education levels, and 20 occupations, then our planning model will have 100,000 viewer segments. Computationally efficient algorithms are highly demanded for real implementations of our framework.

(4) We test the approach on random test cases and two real data sets. An efficient ad serving procedure is designed to incorporate additional constraints such as frequency caps and exclusivity. We show that our approximation of chance constraints is very tight. We explore the trade-off between short-term spot market revenue and the effectiveness of guaranteed ads through numerical experiments. The proposed integrated planning of guaranteed and non-guaranteed ads is shown to be capable of achieving higher profit and also ensuring effective delivery of guaranteed ads.

The remainder of this paper is organized as follows. We first review the literature in the next section. Section 3 then formally defines the problem, presents the DRCC model and then safely approximates the model by using a robust optimization model. Section 4 transforms the robust optimization model into a linear program and quantifies the transformation performance. Section 5 presents the audience clustering algorithm. Section 6 defines the ad serving procedure and presents the numerical study. The paper is concluded in Section 7. All the proofs are provided in the e-companion.

## 2. Literature Review

Ad delivery is a central problem for publishers and has attracted a substantial amount of attention from researchers and practitioners. Particularly, we have witnessed increasing interest in ad delivery from the operations research community. In the literature, however, guaranteed and non-guaranteed ads are often treated separately. Different application contexts have also led to different models and perspectives.

The first stream of research is related to the delivery problem for guaranteed ads. Vee et al. (2010) proposed an online algorithm for the ad delivery of guaranteed advertising by assuming that they have access to random samples from the future set of user arrivals. Based on this algorithm, Bharadwaj et al. (2012) focused on improving the efficiency of the solution procedure, with the objective of balancing the feasibility and representativeness of the solution. Representativeness measures the distance between an ad delivery plan and an ideal plan. Note that an ad campaign targets a set of viewer segments and that the advertiser prefers that its campaign reaches viewers from all segments; hence, an ideal plan assigns each ad campaign equal proportions of impressions from all of its targeted viewer segments (Yang et al. 2010, McAfee et al. 2013). Both of the above works are online algorithms and do not consider the distribution information of the uncertain factors. As in Vee et al. (2010) and Bharadwaj et al. (2012), Turner (2012) also minimized a convex objective function, albeit with a different intention. Specifically, his objective function minimizes the L2 distance between the delivery plan and the ideal representative plan. Surprisingly, he was able to show that under certain conditions, the plan can simultaneously maximize reach and minimize the variance of the number of allocated impressions. Furthermore, he proposed very efficient algorithms to handle large-scale cases, and our algorithm borrows ideas from these algorithms. Deza et al. (2015) formulated the planning problem for guaranteed contracts as a chance-constrained optimization model and solved the model using sample approximation. In comparison, our chance-constrained model is distributionally robust. Hojjat et al. (2017) considered the ad delivery problem for a new type of guaranteed contract. Each contract specifies the reach and a minimum frequency, where frequency is the number of times that each individual is exposed. They developed a flexible framework that aims for minimal under-delivery and proper representativeness. Lejeune and Turner (2019) presented an ad delivery planning method that maximizes the spreading of impressions (across viewer segments and time) measured by the Gini coefficient, while it penalizes the unfulfillment of guaranteed impressions for ad campaigns.

Although all of the above works aimed to achieve robustness, with the exception of Deza et al. (2015), the majority of them did not achieve robustness in a quantifiable manner. In the present work, we explicitly handle the uncertainty using a distributionally robust chance-constrained program.

The second stream of literature is related to the issues of how a publisher should optimize its revenue from non-guaranteed ads. The majority of the research, however, focuses on the auction mechanism design (Edelman et al. 2007, Feng et al. 2007, Chen et al. 2009, Zhu and Wilbur 2011, Abhishek and Hosanagar 2013, Chen 2017). Walsh et al. (2010) suggested adopting an inventory planning approach rather than auction mechanisms. They designed efficient algorithms to compute the optimal ad inventory allocation solution for the scenarios in which the number of user segments is enormous. Najafi-Asadolahi and Fridgeirsdottir (2014) studied the optimal pricing decisions for an online advertising publisher that faces dynamic, price-dependent advertiser arrivals and needs to decide its pricing quotation (“ask price”) to ad networks. Pricing functions as an indirect tool for allocating advertising resources to advertisers. Shamsi et al. (2014) also studied the allocation of non-guaranteed ads but adopted an online planning approach. They proposed a framework for allocating impressions to advertisers with limited budgets to minimize the risk of the ad network.

The above works indicate an increasing interest in using a planning approach rather than auctions for non-guaranteed ads. Note that the so-called planning approach can actually be considered a generalized auction. However, the generalized auction is market-share based (Chen et al. 2009) rather than event based. In other words, the traditional auction implies holding an auction for each event, namely, whenever an ad display opportunity arises. An undesirable consequence is that ads with higher bids may exhaust their daily budget in the early part of a day, resulting in ads with lower bids being shown throughout the rest of the day (Lu et al. 2015). Such a consequence first means non-smooth delivery for the advertiser. It also means a loss of potential revenue due to a lack of global optimization (Benisch et al. 2009). In practice, the publisher and advertisers may address these drawbacks using some ad hoc arrangements, such as intentionally lowering bid prices for certain periods of a day or randomly participating in some of the auctions throughout the day. With the planning approach, ads with higher bids will eventually gain larger shares over a period, although they do not necessarily win each event-based auction. The above drawbacks of the auction mechanism can be greatly alleviated.

The third research stream, which is also the closest to the present study, considers the ad resource allocation problem in the presence of both guaranteed and non-guaranteed ads. Using a stylized model, Araman and Popescu (2010) studied the trade-off between upfront and spot market sales. Although the setting is more tailored to traditional media advertising such as cable networks and TV, their results are of great value to online display advertising. One of their major findings is that "ignoring audience uncertainty can have a significant cost for media capacity planning and allocation". Ghosh et al. (2009) argued that bids signal the value of an ad slot and thus suggested that publishers assign randomized bids (random variables selected from a uniform distribution) to the guaranteed ads and have them compete with the non-guaranteed ads based on bid values. Yang et al. (2010) proposed a multi-objective optimization model for inventory allocation between guaranteed and non-guaranteed ads; however, this model does not consider the uncertainties embedded in the problem. Considering uncertainty, Balseiro et al. (2014) approached the problem of jointly optimizing the quality of guaranteed ads, as measured by the click-through rates (CTRs), and the spot-market revenue. They proposed a bid-price control policy that is similar in spirit to the one proposed by Ghosh et al. (2009). Their dynamic programming approach is more applicable to small- to medium-sized publishers, and they do not consider reach.

To obtain an ad delivery plan that is "quantifiably" robust to the uncertainty of the impression supply, we formulate this ad delivery problem as a DRCC model. Chance-constrained programming was first introduced by Charnes and Cooper (1959) and further strengthened by Miller and Wagner (1965). Many researchers have proposed safe approximation methods to replace chance constraints with tractable deterministic convex constraints. For example, Nemirovski and Shapiro (2006) presented a convex Bernstein approximation of chance-constrained programs, Chen et al. (2007) introduced the forward and backward deviations to approximate chance constraints, Chen et al. (2010) and Zymler et al. (2013) presented the conditional-value-at-risk (CVaR)-based approximations for chance constraints, and Postek et al. (2018) constructed safe approximations of chance constraints by deriving an upper bound on the moment-generating functions of random variables. However, some chance constraints in our model are not affine in either random variables or decision variables, which violates the settings of chance constraints in the above works and hence prevents the adoption of their results. Moreover, in our case, a major challenge is to ensure that the transformed model is practically tractable for large-scale implementation. For this purpose, we will build upon the existing studies and design some new approximation techniques.

# 3. Model

## 3.1. Problem Definition

Consider an online advertising publisher that is planning its ad delivery for a single period, e.g., a day. The publisher receives a large number of user visits to its website, which creates advertising opportunities. The website viewers (the audience) can be categorized into multiple viewer segments according to their characteristics, such as residence and demographics. We assume here that a viewer segment is the finest, that is, it cannot be further divided; thus, a viewer can be categorized into a unique viewer segment only. Note that the arrival of a single viewer may create multiple impressions and that multiple ads can be displayed on the same web page. The supply of ad resources (impressions) is random due to the uncertainties related to viewer arrivals and visit durations. Note that we do not need assumptions on the stationarity of viewer arrivals, which typically show strong of-day and day-of-the-week effects. Rather, we only need information on the aggregated amount of supply in the period.

The publisher serves two types of ads. The guaranteed ads are from contracts that the publisher signed with advertisers in the upfront market. Each guaranteed ad (or ad campaign) is associated with a set of targeted viewer segments and a certain number of impressions to be delivered. The publisher guarantees the advertisers that the contracts will be fulfilled. In the spot market, advertisers submit their bids for each ad, specifying targeted viewer segments, their willingness to pay, and a daily budget. In a period, the revenue from guaranteed ads is fixed, and the publisher can earn additional revenue from the spot market only. However, to ensure advertisers' satisfaction, the publisher will first endeavor to ensure that the contracted numbers of impressions are delivered. Some advertisers also specify a targeted reach for their ad campaigns. Even if no reach is specified for some contracts, to improve advertising effectiveness and to attract more advertisers, the publisher attempts to improve the reach of ad delivery whenever possible. See Section 6.1 for more discussion on reach, and for now, we simply assume that a targeted reach is specified for each guaranteed contract. The advertising resources are limited, perishable, and shared among all advertisements; thus, the publisher faces a trade-off in its ad delivery planning between earning a short-term revenue and maintaining its long-term benefit.

In this paper, we propose an integrated model that allows the publisher to plan its ad delivery for guaranteed ads and non-guaranteed ads simultaneously. The proposed model will consider the uncertainties embedded in the problem and also address the trade-off mentioned earlier. To the best of our knowledge, this is the first work in the literature to make such an attempt.

**Table 1** Notation definitions

| Symbol | Description |
|---|---|
| $\mathcal{V}$ | set of viewer segments |
| $\mathcal{J}$ | set of guaranteed ad campaigns |
| $\mathcal{W}$ | set of non-guaranteed ad campaigns |
| $\mathcal{V}_j$ | set of viewer segments targeted by guaranteed ad campaign $j$ |
| $\mathcal{V}_w$ | set of viewer segments targeted by non-guaranteed ad campaign $w$ |
| $\mathcal{J}_v$ | set of guaranteed ad campaigns that target viewer segment $v$ |
| $\mathcal{W}_v$ | set of non-guaranteed ad campaigns that target viewer segment $v$ |
| $D_j$ | demand of guaranteed ad campaign $j$ |
| $R_j$ | minimum expected reach of guaranteed ad campaign $j$ |
| $B_w$ | budget of non-guaranteed ad campaign $w$ |
| $N_v$ | population size of viewer segment $v$ |
| $\varphi$ | number of available ad slots on a web page |
| $e_{vw}$ | average revenue per impression from non-guaranteed ad campaign $w$ on viewer segment $v$ |
| $\tilde{s}_v$ | supply (impressions) of viewer segment $v$ |
| $\tilde{r}_{vj}$ | reach of guaranteed ad campaign $j$ on viewer segment $v$ |
| **Decision variables:** | |
| $p_{vj}$ | proportion of impressions from viewer segment $v$ allocated to guaranteed ad campaign $j$ |
| $p_{vw}$ | proportion of impressions from viewer segment $v$ allocated to non-guaranteed ad campaign $w$ |

**Notations**. For easy reference, we list the primary notations used in this paper in Table 1. We label random variables with tildes (e.g., the supply of viewer segment $v$, $\tilde{s}_v$, is random), while their realizations are represented by the same symbols without tildes. We also use lower-case boldface letters and upper-case calligraphic letters to denote vectors and sets, respectively. In particular, we define $\boldsymbol{p} = \{p_{vj} \mid v \in \mathcal{V}, j \in \mathcal{J}_v;\ p_{vw} \mid v \in \mathcal{V}, w \in \mathcal{W}_v\}$ as the ad delivery plan, where $p_{vj}$ and $p_{vw}$ denote the proportions of viewer segment $v$'s total impressions allocated to the guaranteed ad campaign $j$ and the non-guaranteed ad campaign $w$, respectively.

Before proceeding to the model, we first list two important assumptions.

ASSUMPTION 1. *Individuals in the same viewer segment are homogeneous. Therefore, an impression from a viewer segment is equally likely to be supplied by any individual in this segment.*

The homogeneity assumption is reasonable because we assume that individuals within a viewer segment share common behavioral patterns. We can always further divide a segment into finer viewer segments if the assumption fails to hold.

ASSUMPTION 2. *The supplies $\{\tilde{s}_v \mid v \in \mathcal{V}\}$ are independent across all viewer segments. Moreover, for each viewer segment $v \in \mathcal{V}$, $\tilde{s}_v$ is symmetrically distributed in $[\mu_v - \hat{s}_v, \mu_v + \hat{s}_v]$ with $\mathbb{E}[\tilde{s}_v] = \mu_v$.*

The assumption essentially says that the users of different segments do not affect each other. We validate the assumption in the e-companion Section EC.1. To explore the effect of the assumption failing to hold, we also test a model without this assumption in the e-companion Section EC.12.

Let $\tilde{\boldsymbol{s}} = \{\tilde{s}_v \mid v \in \mathcal{V}\}$ be a vector of supplies from all viewer segments, and

$$\Xi_{\tilde{\boldsymbol{s}}} = \left\{\tilde{\boldsymbol{s}} \in \mathbb{R}^{|\mathcal{V}|} \mid \mu_v - \hat{s}_v \le \tilde{s}_v \le \mu_v + \hat{s}_v,\ \forall v \in \mathcal{V}\right\}$$

be the support of $\tilde{\boldsymbol{s}}$. Denote the probability distribution of $\tilde{\boldsymbol{s}}$ by $\mathbb{P}_{\tilde{\boldsymbol{s}}}$; based on Assumption 2, $\mathbb{P}_{\tilde{\boldsymbol{s}}}$ belongs to a confidence set of distributions $\mathcal{D}_{\tilde{\boldsymbol{s}}}$ (also called the ambiguity set) defined as follows:

$$\mathcal{D}_{\tilde{\boldsymbol{s}}} = \left\{\mathbb{P}_{\tilde{\boldsymbol{s}}} \in \mathcal{M}(\Xi_{\tilde{\boldsymbol{s}}}) \;\middle|\; \int_{\Xi_{\tilde{\boldsymbol{s}}}} \mathrm{d}\mathbb{P}_{\tilde{\boldsymbol{s}}} = 1,\ \int_{\Xi_{\tilde{\boldsymbol{s}}}} \tilde{\boldsymbol{s}}\ \mathrm{d}\mathbb{P}_{\tilde{\boldsymbol{s}}} = \boldsymbol{\mu}\right\},$$

where $\mathcal{M}(\Xi_{\tilde{\boldsymbol{s}}})$ denotes the set of all symmetric probability distributions on $\Xi_{\tilde{\boldsymbol{s}}}$ that satisfy mutual independence of $\tilde{\boldsymbol{s}}$, and $\boldsymbol{\mu} = \{\mu_v \mid v \in \mathcal{V}\}$.

### 3.2. Reach of the Ad Campaign

To construct our model, we first need to quantify the reach, and for an ad campaign, reach is defined as the total number of unique individuals who have seen the ads in this campaign. A high reach is important for the effectiveness of display advertising and is therefore preferred by advertisers. More than one ad may be used in an ad campaign. However, for our purposes, we do not need to differentiate the ads in the same campaign.

Before an ad delivery plan is realized, the reach of an ad campaign is uncertain for two reasons. (1) The supply process of ad resources is stochastic. It is impossible to know whether an individual will arrive, the number of times that he/she will arrive, and the duration of each visit. (2) Ad serving is subject to randomness, as is the serving result of any ad delivery plan. The ad delivery plan is ex ante, and it specifies only the proportions of viewer segment's impressions allocated to different ad campaigns. When an ad display opportunity arises, a qualified ad is randomly selected according to the plan. We will detail the ad serving procedure in Section 6.2.

Consequently, the reach remains uncertain even if the supply of impressions is known. Confronted with this issue, we measure the expected reach with respect to the random supply of impressions.

PROPOSITION 1 **(Quantifying Reach)**. *The expectation of $\tilde{r}_{vj}$ (reach of the guaranteed ad campaign $j$ on viewer segment $v$) with respect to supply $\tilde{s}_v$ and impression proportion $p_{vj}$ is*

$$\mathbb{E}\left[\tilde{r}_{vj} \mid (\tilde{s}_v, p_{vj})\right] = N_v - N_v \left[1 - \frac{1}{N_v} + \frac{(1-p_{vj})^{\varphi}}{N_v}\right]^{\tilde{s}_v}, \forall j \in \mathcal{J}, v \in \mathcal{V}_j,$$

*and the total expected reach of the guaranteed ad campaign $j$ with respect to supplies $\{\tilde{s}_v \mid v \in \mathcal{V}_j\}$ and impression proportions $\{p_{vj} \mid v \in \mathcal{V}_j\}$ is $\sum_{v \in \mathcal{V}_j} \mathbb{E}[\tilde{r}_{vj} \mid (\tilde{s}_v, p_{vj})]$, $\forall j \in \mathcal{J}$.*

### 3.3. Model Formulation

We are now ready to present our model. Our objective is to find an integrated ad delivery plan that maximizes the total revenue from the spot market while achieving the guaranteed delivery and reach for the upfront market. We formulate the problem into a DRCC program. The model, denoted by ADP-0 (ADP for ad delivery planning), is presented first, followed by detailed explanations.

$$[\text{ADP-0}] \quad \max\ t, \tag{1}$$

$$\text{s.t.} \inf_{\mathbb{P}_{\tilde{s}} \in \mathcal{D}_{\tilde{s}}} \mathbb{P}_{\tilde{s}} \left\{ \sum_{w \in \mathcal{W}} \min\Big( \sum_{v \in \mathcal{V}_w} \varphi \tilde{s}_v e_{vw} p_{vw}, B_w \Big) \geq t \right\} \geq 1 - \varepsilon, \quad \text{(revenue constraint)} \tag{2}$$

$$\inf_{\mathbb{P}_{\tilde{s}} \in \mathcal{D}_{\tilde{s}}} \mathbb{P}_{\tilde{s}} \left\{ \sum_{v \in \mathcal{V}_j} \varphi \tilde{s}_v p_{vj} \geq D_j \right\} \geq 1 - \varepsilon_j^d, \ \forall j \in \mathcal{J}, \quad \text{(delivery constraint)} \tag{3}$$

$$\inf_{\mathbb{P}_{\tilde{s}} \in \mathcal{D}_{\tilde{s}}} \mathbb{P}_{\tilde{s}} \left\{ \sum_{v \in \mathcal{V}_j} \mathbb{E}\left[\tilde{r}_{vj} \mid (\tilde{s}_v, p_{vj})\right] \geq R_j \right\} \geq 1 - \varepsilon_j^r, \ \forall j \in \mathcal{J}, \quad \text{(reach constraint)} \tag{4}$$

$$\sum_{j \in \mathcal{J}_v} p_{vj} + \sum_{w \in \mathcal{W}_v} p_{vw} \leq 1, \ \forall v \in \mathcal{V}, \quad \text{(supply constraint)} \tag{5}$$

$$0 \leq p_{vj} \leq 1, \ \forall v \in \mathcal{V},\ j \in \mathcal{J}_v; \ 0 \leq p_{vw} \leq 1, \ \forall v \in \mathcal{V},\ w \in \mathcal{W}_v. \tag{6}$$

**Objective Function**. Because there are $\varphi$ available ad slots on a webpage ($\varphi \geq 1$), the number of impressions from viewer segment $v$ is $\varphi \tilde{s}_v$, and the impressions allocated to non-guaranteed ad campaign $w \in \mathcal{W}_v$ in the spot market is $\varphi \tilde{s}_v p_{vw}$. The average revenue per impression from ad campaign $w$ on segment $v$ is $e_{vw}$. Note that adopting our approach implies that the publisher is not running event-based auctions, and the bids help advertisers win a share of the targeted ad resources. For $e_{vw}$, it is simply the bid price if the advertiser does not adjust its bid in the period, which is often the case; otherwise, it can be estimated based on the eCPM (effective cost per mille impression) of this ad campaign or the most similar ad campaigns. Given that the budget of the non-guaranteed ad campaign cannot be over-consumed, the revenue from ad campaign $w$ is $\min\{\sum_{v \in \mathcal{V}_w} \varphi \tilde{s}_v e_{vw} p_{vw}, B_w\}$, which is a truncated random variable (with uncertainty from the impression supply) and is difficult to model using traditional approaches (e.g., maximizing the expectation). Therefore, we seek to maximize the $\varepsilon-$percentile of the total revenue from non-guaranteed campaigns, which equals $t$ as stipulated by (1) and (2). Furthermore, maximizing the percentile also allows the publisher to incorporate its risk attitude by adjusting the value of $\varepsilon$.

**Delivery and Reach**. Constraint (3) guarantees that the demand of each ad campaign $j$ is satisfied with a probability of no less than $(1 - \varepsilon_j^d)$ and constraint (4) ensures that the expected reach of ad campaign $j$ is at least $R_j$ with a probability of no less than $(1 - \varepsilon_j^r)$. We will discuss how to set the risk levels $(\varepsilon, \varepsilon_j^d, \varepsilon_j^r)$ in Section 6.3 and for now, we assume they are exogenously given.

**Other Constraints**. Constraint (5) prevents the total advertising resources of each viewer segment $v$ from being over-allocated, and constraint (6) maintains the validity of decision variables $p_{vj}$ and $p_{vw}$. For convenience, in the following, constraints (2), (3), (4), and (5) are called the *revenue*, *delivery*, *reach*, and *supply* constraints, respectively.

### 3.4. Safe Approximation

ADP-0 is difficult to solve because (1) it is not in closed form and can be non-convex (Nemirovski and Shapiro 2006), and (2) the reach constraint is not affine in either decision variables or random variables, thereby preventing the direct application of recently developed methods for chance-constrained programs (e.g., Zymler et al. 2013, Postek et al. 2018). In robust optimization, Ben-Tal et al. (2015) used Fenchel duality to construct tractable robust counterparts of nonlinear uncertain inequalities, which involves calculating the conjugate function of the constraint function. However, in our model, calculating the conjugate function of the reach constraint function is complex and time consuming; therefore, their result is not suitable for our case.

In this section, we transform ADP-0 into a more tractable robust optimization model via safe approximation of the chance constraints. Considering the nature of our problem, the tractability of the resulting model is critical in choosing the approximation approach. Li et al. (2012) compared methods for approximating chance constraints with different uncertainty sets, including box, ellipsoid, polyhedron, “interval + ellipsoid”, and “interval + polyhedron”. They showed that all five approximation approaches share similar bounds on the probability of constraint violation. We choose to work with the “interval + polyhedron” uncertainty set, which leads to a more tractable robust model. In particular, the following probability bound derived by Bertsimas and Sim (2004) is used, and they have shown that the bound is tighter than the other classical bounds in the literature, particularly when the number of random variables is large.

THEOREM 1 **(Violation Probability Bound, Bertsimas and Sim (2004))**. *Let $\{\tilde{\xi}_k \mid k \in \mathcal{K}\}$ be a set of $n$ independent random variables, where for each $k \in \mathcal{K}$, $\tilde{\xi}_k$ is symmetrically distributed in the interval* $[-1,1]$. *Given a set of bounded coefficients $\{\lambda_k \in [0,1] \mid k \in \mathcal{K}\}$, for any scalar $\Gamma \in [1,n]$, we have*

$$\mathbb{P}_{\tilde{\xi}}\left\{\sum_{k\in\mathcal{K}} \lambda_k \tilde{\xi}_k \geq \Gamma\right\} \leq g(n,\Gamma),$$

*where*

$$g(n,\Gamma) = \frac{1}{2^n}\left[\left(1 - \frac{\Gamma+n}{2} + \left\lfloor\frac{\Gamma+n}{2}\right\rfloor\right)\binom{n}{\lfloor\frac{\Gamma+n}{2}\rfloor} + \sum_{m=\lfloor\frac{\Gamma+n}{2}\rfloor+1}^{n}\binom{n}{m}\right].$$

As shown by Bertsimas and Sim (2004), $g(n,\Gamma)$ in Theorem 1 is the best possible bound, i.e., it is achievable. Accordingly, our approximations are tight, as our experiments in Section 6.4 will show.

LEMMA 1. *The function $g(n,\Gamma)$, where $\Gamma \in [1,n]$ and $n$ is a positive integer, is strictly decreasing in $\Gamma$; and moreover, $g(n,1) = 1/2$, $g(n,n) = 1/2^n$.*

By Lemma 1, for a given $n$, as $\Gamma$ ranges from 1 to $n$, $g(n,\Gamma)$ is strictly decreasing from $1/2$ to $1/2^n$. In our approximations, we need to connect the $\Gamma$ with the risk level (i.e., $\varepsilon$, $\varepsilon_j^d$, $\varepsilon_j^r$) via this function. To this end, we define the following function that outputs a minimum $\Gamma$ that guarantees the violation probability.

$$g^{-1}(n,\varepsilon) = \begin{cases} \min\limits_{\Gamma\in[1,n]\,:\,g(n,\Gamma)\leq\varepsilon} \Gamma, & \text{if } \varepsilon \geq 1/2^n, \\ n, & \text{if } \varepsilon < 1/2^n. \end{cases}$$

PROPOSITION 2 **(Delivery Approximation)**. *A solution $\{p_{vj}^* \mid v \in \mathcal{V}_j, j \in \mathcal{J}\}$ must fulfill the delivery constraint (3) in ADP-0 if it satisfies the constraint*

$$\min_{\boldsymbol{\theta}_j\in\mathcal{F}_{\boldsymbol{\theta}_j}(\Gamma_j^d)} \left\{\sum_{v\in\mathcal{V}_j} \varphi\left(\mu_v + \hat{s}_v\theta_{vj}\right) p_{vj}\right\} \geq D_j,\ \forall j \in \mathcal{J},$$

*where $\Gamma_j^d = g^{-1}(|\mathcal{V}_j|, \varepsilon_j^d)$ and*

$$\mathcal{F}_{\boldsymbol{\theta}_j}(\Gamma_j^d) = \left\{\theta_{vj},\ v\in\mathcal{V}_j \;\middle|\; \begin{array}{l} \sum_{v\in\mathcal{V}_j}\theta_{vj} \geq -\Gamma_j^d \\ -1 \leq \theta_{vj} \leq 0,\ \forall v \in \mathcal{V}_j \end{array}\right\},\ \forall j \in \mathcal{J}.$$

To approximate the reach constraint in ADP-0, we first approximate the expected reach $\mathbb{E}[\tilde{r}_{vj} \mid \tilde{s}_v, p_{vj}]$ by its lower bound that is bilinear in $\tilde{s}_v$ and $p_{vj}$, as stated in the following lemma.

Lemma 2. *For any $v \in \mathcal{V}_j, j \in \mathcal{J}$, the expected reach $\mathbb{E}[\tilde{r}_{vj} \mid (\tilde{s}_v, p_{vj})]$ has a lower bound:*

$$\min_{l \in \mathcal{L}} \left\{ \alpha_v^l \tilde{s}_v p_{vj} + \beta_v^l \tilde{s}_v + \gamma_v^l p_{vj} + \delta_v^l \right\},$$

*where $\mathcal{L} = \{1, \ldots, L-1\}$ is a predefined set of indices ($L \geq 2$) and $\{\alpha_v^l, \beta_v^l, \gamma_v^l, \delta_v^l \mid l \in \mathcal{L}\}$ is a set of parameters defined as*

$$\begin{cases} \alpha_v^l = [f_v^1(\hat{p}_v^{l+1}) - f_v^1(\hat{p}_v^l)] \,/\, (\hat{p}_v^{l+1} - \hat{p}_v^l), \;\; \beta_v^l = [\hat{p}_v^{l+1} f_v^1(\hat{p}_v^l) - \hat{p}_v^l f_v^1(\hat{p}_v^{l+1})] \,/\, (\hat{p}_v^{l+1} - \hat{p}_v^l), \\ \gamma_v^l = [f_v^2(\hat{p}_v^{l+1}) - f_v^2(\hat{p}_v^l)] \,/\, (\hat{p}_v^{l+1} - \hat{p}_v^l), \;\; \delta_v^l = [\hat{p}_v^{l+1} f_v^2(\hat{p}_v^l) - \hat{p}_v^l f_v^2(\hat{p}_v^{l+1})] \,/\, (\hat{p}_v^{l+1} - \hat{p}_v^l), \end{cases}$$

*where $\{\hat{p}_v^l \mid l = 1, \ldots, L\}$ is a set of predefined points such that $0 = \hat{p}_v^1 < \cdots < \hat{p}_v^L = 1$, $f_v^1(p)$ and $f_v^2(p)$ are functions defined as*

$$\begin{cases} f_v^1(p) = [f_v(\mu_v + \hat{s}_v, p) - f_v(\mu_v - \hat{s}_v, p)] \,/ 2\hat{s}_v, \\ f_v^2(p) = [(\mu_v + \hat{s}_v) f_v(\mu_v - \hat{s}_v, p) - (\mu_v - \hat{s}_v) f_v(\mu_v + \hat{s}_v, p)] \,/ 2\hat{s}_v, \\ f_v(s, p) = N_v - N_v \left[1 - 1/N_v + (1-p)^{\varphi} / N_v\right]^s. \end{cases}$$

Remark 1. From Proposition 1, it is easy to verify that $\mathbb{E}[\tilde{r}_{vj} | (\tilde{s}_v, p_{vj})]$ is concave in $\tilde{s}_v$ and $p_{vj}$, respectively. The lower bound in Lemma 2 is obtained via two steps: (1) approximating $\mathbb{E}[\tilde{r}_{vj} \mid (\tilde{s}_v, p_{vj})]$ by its secant on $\tilde{s}_v$ within the interval $[\mu_v - \hat{s}_v, \mu_v + \hat{s}_v]$, denoted by $F(\tilde{s}_v, p_{vj})$, which is linear in $\tilde{s}_v$ and nonlinear in $p_{vj}$; and (2) approximating $F(\tilde{s}_v, p_{vj})$ by its piecewise linear interpolation on $p_{vj}$ with breakpoints $\{\hat{p}_v^l \mid l = 1, \ldots, L\}$ in the interval $[0, 1]$. Note that in the second step, given $L$, we fix $\tilde{s}_v$ as its mean value $\mu_v$, and follow the scheme of Kontogiorgis (2000) to automatically select the breakpoints. Specifically, $\hat{p}_v^1 = 0$, $\hat{p}_v^L = 1$, and $\hat{p}_v^l$ is the solution to the following equation:

$$\int_0^{\hat{p}_v^l} \left| \frac{\mathrm{d}^2 F(\mu_v, p)}{\mathrm{d}p^2} \right|^{\frac{1}{2}} \mathrm{d}p = \frac{l-1}{L-1} \int_0^1 \left| \frac{\mathrm{d}^2 F(\mu_v, p)}{\mathrm{d}p^2} \right|^{\frac{1}{2}} \mathrm{d}p, \; l = 2, \ldots, L-1.$$

Kontogiorgis (2000) has shown that under the above scheme, the approximation error between $F(\mu_v, p_{vj})$ and its piecewise linear interpolation will decrease at the optimal rate of $O(L^{-2})$.

Proposition 3 **(Reach Approximation)**. *A solution $\{p_{vj}^* \mid v \in \mathcal{V}_j, j \in \mathcal{J}\}$ must fulfill the reach constraint (4) in ADP-0 if it satisfies the following constraint*

$$\min_{\eta_j \in \mathcal{F}_{\eta_j}(\Gamma_j^r)} \left\{ \sum_{v \in \mathcal{V}_j} \min_{l \in \mathcal{L}} \left\{ a_v^l(p_{vj}) \eta_{vj} + b_v^l(p_{vj}) \right\} \right\} \geq R_j, \; \forall j \in \mathcal{J},$$

*where* $\Gamma_j^r = g^{-1}(|\mathcal{V}_j|, \varepsilon_j^r)$ *and*

$$\mathcal{F}_{\eta_j}(\Gamma_j^r) = \left\{ \eta_{vj},\, v \in \mathcal{V}_j \;\middle|\; \begin{array}{l} \sum_{v\in\mathcal{V}_j} \eta_{vj} \geq -\Gamma_j^r \\ -1 \leq \eta_{vj} \leq 0,\ \forall v \in \mathcal{V}_j \end{array} \right\}, \forall j \in \mathcal{J},$$

$$a_v^l(p_{vj}) = \hat{s}_v \alpha_v^l p_{vj} + \hat{s}_v \beta_v^l,\ \forall v \in \mathcal{V}_j, l \in \mathcal{L},$$

$$b_v^l(p_{vj}) = \left(\mu_v \alpha_v^l + \gamma_v^l\right) p_{vj} + \mu_v \beta_v^l + \delta_v^l,\ \forall v \in \mathcal{V}_j, l \in \mathcal{L}.$$

REMARK 2. The probability terms in delivery and reach constraints are replaced by deterministic constraints as stipulated in Propositions 2 and 3. This is achieved via approximating the probability terms with their deterministic lower bounds (obtained by Theorem 1).

Motivated by the above approximations, we approximate the revenue constraint and objective function by the following new objective function.

$$\max_{\boldsymbol{p}} \min_{\boldsymbol{\phi}\in\mathcal{F}_{\boldsymbol{\phi}}(\Gamma)} \left\{ \sum_{w\in\mathcal{W}} \min\left\{ \sum_{v\in\mathcal{V}_w} \varphi e_{vw} \left(\mu_v + \hat{s}_v \phi_v\right) p_{vw}, B_w \right\} \right\},$$

where $\Gamma = g^{-1}(|\mathcal{V}|, \varepsilon)$, and

$$\mathcal{F}_{\boldsymbol{\phi}}(\Gamma) = \left\{ \phi_v,\, v \in \mathcal{V} \;\middle|\; \begin{array}{l} \sum_{v\in\mathcal{V}} \phi_v \geq -\Gamma \\ -1 \leq \phi_v \leq 0,\ \forall v \in \mathcal{V} \end{array} \right\}.$$

Subsequently, we arrive at the following deterministic model, denoted by ADP-1.

$$\text{[ADP-1]} \quad \max_{\boldsymbol{p}} \min_{\boldsymbol{\phi}\in\mathcal{F}_{\boldsymbol{\phi}}(\Gamma)} \sum_{w\in\mathcal{W}} \min\left\{ \sum_{v\in\mathcal{V}_w} \varphi e_{vw} \left(\mu_v + \hat{s}_v \phi_v\right) p_{vw}, B_w \right\} \quad \text{(revenue objective)} \tag{7}$$

$$\text{s.t.} \quad \min_{\boldsymbol{\theta}_j\in\mathcal{F}_{\boldsymbol{\theta}_j}(\Gamma_j^d)} \sum_{v\in\mathcal{V}_j} \varphi\left(\mu_v + \hat{s}_v \theta_{vj}\right) p_{vj} \geq D_j,\ \forall j \in \mathcal{J}, \quad \text{(delivery constraint)} \tag{8}$$

$$\min_{\boldsymbol{\eta}_j\in\mathcal{F}_{\boldsymbol{\eta}_j}(\Gamma_j^r)} \sum_{v\in\mathcal{V}_j} \min_{l\in\mathcal{L}} \left\{ a_v^l(p_{vj})\eta_{vj} + b_v^l(p_{vj}) \right\} \geq R_j,\ \forall j \in \mathcal{J}, \quad \text{(reach constraint)} \tag{9}$$

constraints (5) and (6).

Note that the decision variable $t$ is not present in model ADP-1. In fact, $t$ is not a decision, and there is no feasibility issue as well for the revenue constraint (2): many feasible $t$ values are available for any given $\boldsymbol{p}$. As specified in the following theorem, we can employ a semidefinite programming (SDP) model to obtain the best $t$ for a given $\boldsymbol{p}$, denoted by $t^*$. Note that the SDP model can be solved easily thanks to its limited scale, which is largely determined by the number of campaigns, i.e., $|\mathcal{W}|$. Theorem 2 says that model ADP-1 effectively provides a safe approximation of ADP-0 indirectly.

THEOREM 2 **(Safe Approximation)**. *Let $\boldsymbol{p}^*$ be an optimal solution to ADP-1. Based on $\boldsymbol{p}^*$, we can obtain a $t^*$ such that $(\boldsymbol{p}^*, t^*)$ is a feasible solution to ADP-0, where*

$$t^* = \sum_{w\in\mathcal{W}} \sum_{v\in\mathcal{V}_w} \varphi e_{vw} \mu_v p^*_{vw} - \frac{T\left(\boldsymbol{p}^*\right)}{\varepsilon},$$

*and $T(\boldsymbol{p}^*)$ is the optimal objective value of the following model.*

$$\begin{aligned}
&\min \ \varepsilon z + \psi + \langle \boldsymbol{\Sigma}, \boldsymbol{X} \rangle, \\
&s.t. \begin{bmatrix} \boldsymbol{X}, & \boldsymbol{A}_1 \\ \\ \boldsymbol{A}_1^\intercal, & M_1 \end{bmatrix} \succeq 0, \begin{bmatrix} 2\boldsymbol{X}, & \boldsymbol{A}_2 \\ \\ \boldsymbol{A}_2^\intercal, & M_2 \end{bmatrix} \succeq 0, \begin{bmatrix} 2\boldsymbol{X}, & \boldsymbol{A}_3 \\ \\ \boldsymbol{A}_3^\intercal, & M_3 \end{bmatrix} \succeq 0, \\
&\qquad \boldsymbol{X} \in \mathbb{S}^{|\mathcal{W}|} \text{ and } \boldsymbol{X} \succeq 0,
\end{aligned}$$

*where $\boldsymbol{\Sigma} \in \mathbb{S}^{|\mathcal{W}|}$ is a symmetric matrix of dimension $|\mathcal{W}|$ defined by*

$$\Sigma_{w_1,w_2} = \sum_{v\in(\mathcal{V}_{w_1}\cap\mathcal{V}_{w_2})} \varphi^2 e_{vw_1} e_{vw_2} \hat{s}_v^2 p^*_{vw_1} p^*_{vw_2}, \forall w_1, w_2 \in \mathcal{W},$$

*$\langle \boldsymbol{\Sigma}, \boldsymbol{X} \rangle$ is the Frobenius inner product of two matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{X}$, $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \in \mathbb{R}^{|\mathcal{W}|}; M_1, M_2, M_3 \in \mathbb{R}\}$ is defined based on $\boldsymbol{p}^*$ and detailed in the e-companion Section EC.4.*

## 4. Model Transformation

The objective function of ADP-1 is in the form of "max-min-min", and its reach constraint has a "min-min" sub-problem. For problems of this feature, Ardestani-Jaafari and Delage (2016) propose two approximation approaches. Their SDP approximation is not applicable here due to the large scale of the problem. We adopt a linear approximation similar to theirs. However, we extend their results by establishing (1) the exactness of the reach constraint transformation under general uncertainty budgets and (2) an analytical performance bound for the objective function, which is absent in their paper.

### 4.1. Transformation of Objective Function

The objective function (7) in ADP-1 is equivalent to

$$\max_{\boldsymbol{p}} \ \Pi(\boldsymbol{p}),$$

where

$$
\begin{aligned}
\Pi(\boldsymbol{p}) = \min \sum_{w\in\mathcal{W}} & y_w\Big(\sum_{v\in\mathcal{V}_w}\varphi e_{vw}(\mu_v+\hat{s}_v\phi_v)p_{vw}\Big)+\sum_{w\in\mathcal{W}}(1-y_w)B_w,\\
\text{s.t. } & \sum_{v\in\mathcal{V}}\phi_v \geq -\Gamma,\\
& -1\leq \phi_v\leq 0,\ \forall v\in\mathcal{V},\\
& y_w\in\{0,1\},\ \forall w\in\mathcal{W}.
\end{aligned}
$$

In model $\Pi(\boldsymbol{p})$, the variable $y_w$ is a binary such that, for a given $\boldsymbol{\phi}$, $y_w=0$ if the consumption of non-guaranteed campaign $w$ is greater than its budget $B_w$, and $y_w=1$ otherwise.

Note that the objective function in the above model has a set of quadratic terms $\{\phi_v y_w \mid v\in\mathcal{V}_w, w\in\mathcal{W}\}$ and is thus difficult to solve. We linearize the model by introducing a set of auxiliary variables $\{\Delta_{vw}\mid v\in\mathcal{V}_w, w\in\mathcal{W}\}$, and replacing each $\phi_v y_w$ with $\Delta_{vw}$ such that $\Delta_{vw}\geq\max\{\phi_v,-y_w\}$. Subsequently, we have the following lemma.

LEMMA 3. *Model $\Pi(\boldsymbol{p})$ is equivalent to a mixed integer programming model as follows:*

$$
\begin{aligned}
\Pi(\boldsymbol{p}) = \min \sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{V}_w} & \varphi e_{vw}\hat{s}_v p_{vw}\Delta_{vw}+\sum_{w\in\mathcal{W}}\Big(\sum_{v\in\mathcal{V}_w}\varphi e_{vw}\mu_v p_{vw}-B_w\Big)y_w+\sum_{w\in\mathcal{W}}B_w,\\
s.t. & \sum_{v\in\mathcal{V}}\phi_v\geq -\Gamma,\\
& \Delta_{vw}\geq -y_w,\ \forall v\in\mathcal{V}_w,\ w\in\mathcal{W},\\
& \Delta_{vw}\geq \phi_v,\ \forall v\in\mathcal{V}_w,\ w\in\mathcal{W},\\
& \phi_v\leq 0,\ \forall v\in\mathcal{V},\\
& y_w\in\{0,1\},\ \forall w\in\mathcal{W}.
\end{aligned}
$$

We relax the binary constraint on $\{y_w\mid w\in\mathcal{W}\}$, and denote the relaxed linear program by $\hat{\Pi}(\boldsymbol{p})$. The dual of $\hat{\Pi}(\boldsymbol{p})$ can be formulated as (detailed in the e-companion Section EC.5):

$$
\begin{aligned}
\text{dual-}\hat{\Pi}(\boldsymbol{p}) = \max & -\Gamma x+\sum_{w\in\mathcal{W}}h_w,\\
\text{s.t. } & x\geq\sum_{w\in\mathcal{W}_v}m_{vw},\ \forall v\in\mathcal{V},\\
& h_w-\sum_{v\in\mathcal{V}_w}m_{vw}\leq\sum_{v\in\mathcal{V}_w}\varphi e_{vw}(\mu_v-\hat{s}_v)p_{vw},\forall w\in\mathcal{W},
\end{aligned}
$$

$$0 \le m_{vw} \le \varphi e_{vw} \hat{s}_v p_{vw},\ \forall v \in \mathcal{V}_w, w \in \mathcal{W},$$

$$h_w \le B_w,\ \forall w \in \mathcal{W}.$$

Since $\hat{\Pi}(\boldsymbol{p})$ has a finite optimal value, strong duality holds between $\hat{\Pi}(\boldsymbol{p})$ and dual-$\hat{\Pi}(\boldsymbol{p})$. We then have the following proposition.

PROPOSITION 4 **(Objective Transformation)**. *The objective* $\max_{\boldsymbol{p}}$ *dual-*$\hat{\Pi}(\boldsymbol{p})$ *is a lower bound of the objective function (7) in ADP-1.*

REMARK 3. By $\max_{\boldsymbol{p}}$ dual-$\hat{\Pi}(\boldsymbol{p})$, the inner minimization of the “max-min” objective function of $\max_{\boldsymbol{p}} \Pi(\boldsymbol{p})$ is transformed into a maximization problem, which can be further combined with the outer maximization. Since dual-$\hat{\Pi}(\boldsymbol{p}) \le \Pi(\boldsymbol{p})$ for any given plan $\boldsymbol{p}$, there is a potential transformation error. Theoretical error bounds are discussed in Section 4.3.

### 4.2. Transformation of Constraints

The left-hand side of the delivery constraint (8) in ADP-1 involves a minimization problem, which can be replaced by its dual. We then perform an exact transformation of the constraint as follows.

PROPOSITION 5 **(Delivery Transformation)**. *The delivery constraint (8) in ADP-1 can be equivalently transformed into the following set of constraints.*

$$\begin{cases} \sum_{v \in \mathcal{V}_j} \varphi \mu_v p_{vj} - \sum_{v \in \mathcal{V}_j} m^d_{vj} - \Gamma^d_j h^d_j \ge D_j,\ \forall j \in \mathcal{J}, \\ h^d_j + m^d_{vj} \ge \varphi \hat{s}_v p_{vj},\ \forall v \in \mathcal{V}_j, j \in \mathcal{J}, \\ h^d_j \ge 0,\ \forall j \in \mathcal{J}, \\ m^d_{vj} \ge 0,\ \forall v \in \mathcal{V}_j, j \in \mathcal{J}. \end{cases}$$

Next, we transform the reach constraint (9) in ADP-1. By introducing binary variables $\{y^l_{vj} \mid v \in \mathcal{V}_j, l \in \mathcal{L}\}$, the reach constraint is equivalent to

$$Q_j(\boldsymbol{p}) \ge R_j,\ \forall j \in \mathcal{J},$$

where

$$Q_j(\boldsymbol{p}) = \min \sum_{v \in \mathcal{V}_j} \sum_{l \in \mathcal{L}} a^l_v(p_{vj}) y^l_{vj} \eta_{vj} + b^l_v(p_{vj}) y^l_{vj},$$

$$\text{s.t.} \sum_{l \in \mathcal{L}} y^l_{vj} = 1,\ \forall v \in \mathcal{V}_j,$$

$$
\begin{aligned}
&\sum_{v\in\mathcal{V}_j}\eta_{vj}\geq-\Gamma_j^r,\\
&-1\leq\eta_{vj}\leq 0,\ \forall v\in\mathcal{V}_j,\\
&y_{vj}^l\in\{0,1\}\,,\ \forall v\in\mathcal{V}_j, l\in\mathcal{L}.
\end{aligned}
$$

Similar to that in model $\Pi(\boldsymbol{p})$, the objective function of model $Q_j(\boldsymbol{p})$ has a quadratic term $y_{vj}^l\eta_{vj}$. Therefore, we can transform the model into a mixed integer linear program by a similar approach. Specifically, we first introduce auxiliary variables $\{\Delta_{vj}^l \mid v\in\mathcal{V}_j,\ l\in\mathcal{L}\}$ and replace each $y_{vj}^l\eta_{vj}$ by $\Delta_{vj}^l$ with two additional constraints: $-y_{vj}^l\leq\Delta_{vj}^l\leq 0$ and $\sum_{l\in\mathcal{L}}\Delta_{vj}^l=\eta_{vj}$. Fortunately, we are able to prove here that the linear relaxation of the resulting model has no relaxation error, as shown in the following proposition.

THEOREM 3 **(Perfect Formulation)**. *$Q_j(\boldsymbol{p})$ is equivalent to the following linear program.*

$$
\begin{aligned}
\min&\sum_{v\in\mathcal{V}_j}\sum_{l\in\mathcal{L}}a_v^l(p_{vj})\Delta_{vj}^l+b_v^l(p_{vj})y_{vj}^l,\\
s.t.&\sum_{l\in\mathcal{L}}y_{vj}^l=1,\ \forall v\in\mathcal{V}_j,\\
&\sum_{v\in\mathcal{V}_j}\sum_{l\in\mathcal{L}}\Delta_{vj}^l\geq-\Gamma_j^r,\\
&-y_{vj}^l\leq\Delta_{vj}^l\leq 0,\ \forall v\in\mathcal{V}_j, l\in\mathcal{L},\\
&y_{vj}^l\geq 0,\ \forall v\in\mathcal{V}_j, l\in\mathcal{L}.
\end{aligned}
$$

Based on the above theorem and duality theory, we can perform an exact transformation of the reach constraint.

PROPOSITION 6 **(Reach Transformation)**. *The reach constraint (9) in ADP-1 can be equivalently transformed into the following set of constraints.*

$$
\begin{cases}
-\Gamma_j^r h_j^r+\sum_{v\in\mathcal{V}_j}m_{vj}^r\geq R_j,\ \forall j\in\mathcal{J},\\
-h_j^r+m_{vj}^r\leq-a_v^l(p_{vj})+b_v^l(p_{vj}),\ \forall v\in\mathcal{V}_j, j\in\mathcal{J}, l\in\mathcal{L},\\
m_{vj}^r\leq b_v^l(p_{vj}),\ \forall v\in\mathcal{V}_j, j\in\mathcal{J}, l\in\mathcal{L},\\
h_j^r\geq 0,\ \forall j\in\mathcal{J}.
\end{cases}
$$

### 4.3. Performance Analysis of Transformation

Based on the results in Sections 4.1 and 4.2, ADP-1 can be transformed into the following linear program, denoted by ADP-2:

$$\text{[ADP-2]} \ \max -\Gamma x + \sum_{w\in\mathcal{W}} h_w, \tag{10}$$

$$\text{s.t. } x \geq \sum_{w\in\mathcal{W}_v} m_{vw}, \ \forall v\in\mathcal{V}, \tag{11}$$

$$h_w - \sum_{v\in\mathcal{V}_w} m_{vw} \leq \sum_{v\in\mathcal{V}_w} \varphi e_{vw}(\mu_v - \hat{s}_v)p_{vw}, \forall w\in\mathcal{W}, \tag{12}$$

$$m_{vw} \leq \varphi e_{vw}\hat{s}_v p_{vw}, \ \forall v\in\mathcal{V}_w, w\in\mathcal{W}, \tag{13}$$

$$h_w \leq B_w, \ \forall w\in\mathcal{W}, \tag{14}$$

$$\sum_{v\in\mathcal{V}_j} \varphi\mu_v p_{vj} - \sum_{v\in\mathcal{V}_j} m^d_{vj} - \Gamma^d_j h^d_j \geq D_j, \ \forall j\in\mathcal{J}, \tag{15}$$

$$h^d_j + m^d_{vj} \geq \varphi\hat{s}_v p_{vj}, \ \forall v\in\mathcal{V}_j, j\in\mathcal{J}, \tag{16}$$

$$-\Gamma^r_j h^r_j + \sum_{v\in\mathcal{V}_j} m^r_{vj} \geq R_j, \ \forall j\in\mathcal{J}, \tag{17}$$

$$-h^r_j + m^r_{vj} \leq -a^l_v(p_{vj}) + b^l_v(p_{vj}), \ \forall v\in\mathcal{V}_j, j\in\mathcal{J}, l\in\mathcal{L}, \tag{18}$$

$$m^r_{vj} \leq b^l_v(p_{vj}), \ \forall v\in\mathcal{V}_j, j\in\mathcal{J}, l\in\mathcal{L}, \tag{19}$$

$$\sum_{j\in\mathcal{J}_v} p_{vj} + \sum_{w\in\mathcal{W}_v} p_{vw} \leq 1, \ \forall v\in\mathcal{V}, \tag{20}$$

$$0 \leq p_{vj} \leq 1, \ m^d_{vj} \geq 0, \ \forall v\in\mathcal{V}_j, \ j\in\mathcal{J}, \tag{21}$$

$$0 \leq p_{vw} \leq 1, \ m_{vw} \geq 0, \ \forall v\in\mathcal{V}_w, \ w\in\mathcal{W}, \tag{22}$$

$$h^d_j \geq 0, \ h^r_j \geq 0, \ \forall j\in\mathcal{J}. \tag{23}$$

For simplicity, ADP-2 can be rewritten as follows:

$$\text{[ADP-2]} \ \max_{\boldsymbol{p}\in\mathcal{P}} \text{dual-}\hat{\Pi}(\boldsymbol{p}),$$

where $\mathcal{P}$ is a polyhedron defined by $\mathcal{P} = \{\boldsymbol{p} \mid \boldsymbol{p} \text{ satisfies constraints } (15) \sim (23)\}$.

Model ADP-2 is a linear program that can be efficiently solved; therefore, it is considerably more tractable than ADP-1. However, there is an optimality gap between ADP-1 and ADP-2. Denote the optimality gap by

$$\Omega = Z^{ADP-1} - Z^{ADP-2},$$

where $Z^{ADP-1}$ and $Z^{ADP-2}$ are the optimal objective values of ADP-1 and ADP-2, respectively. By Proposition 4, it follows that $Z^{ADP-2} \leq Z^{ADP-1}$; thus, $\Omega \geq 0$. In the following, we aim to find (1) the conditions under which $\Omega = 0$ and (2) an upper bound of $\Omega$.

LEMMA 4. *Let $\boldsymbol{p}^*$ be an optimal solution of ADP-2, and $(\boldsymbol{y}^*, \boldsymbol{\phi}^*)$ be an optimal solution of $\hat{\Pi}(\boldsymbol{p}^*)$. We have $\Omega = 0$ if $y^*_w \in \{0,1\}$ for all $w \in \mathcal{W}$, and $\Omega > 0$ otherwise.*

Lemma 4 provides a trivial ex post condition to check whether the solution to ADP-2 has positive performance gap. In the case that it does, the following optimality condition for $\hat{\Pi}(\boldsymbol{p})$ can be used to analyze the optimality gap $\Omega$.

LEMMA 5. *Given $\boldsymbol{p} \in \mathcal{P}$, denote the optimal solution of $\hat{\Pi}(\boldsymbol{p})$ by $(\boldsymbol{y}^*, \boldsymbol{\phi}^*)$. For each $w \in \mathcal{W}$, we have:*

- *$y^*_w = 0$ if $\sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p_{vw} \geq B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v < 0} \varphi e_{vw} \hat{s}_v p_{vw}$;*
- *$y^*_w = 1$ if $\sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p_{vw} \leq B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v = -1} \varphi e_{vw} \hat{s}_v p_{vw}$;*
- *$0 < y^*_w < 1$ if $B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v = -1} \varphi e_{vw} \hat{s}_v p_{vw} < \sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p_{vw} < B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v < 0} \varphi e_{vw} \hat{s}_v p_{vw}$, and moreover, $y^*_w = -\phi^*_{v'}$, where $v' \in \mathcal{V}_w$ is uniquely determined by the inequality $B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v < \phi^*_{v'}} \varphi e_{vw} \hat{s}_v p_{vw} \leq \sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p_{vw} \leq B_w + \sum_{v \in \mathcal{V}_w : \phi^*_v \leq \phi^*_{v'}} \varphi e_{vw} \hat{s}_v p_{vw}$.*

For a given $\boldsymbol{p} \in \mathcal{P}$, Lemma 5 provides the condition under which the optimal $\boldsymbol{y}^*$ for $\hat{\Pi}(\boldsymbol{p})$ is binary. Since $\hat{\Pi}(\boldsymbol{p})$ is obtained by linearizing the binary variable $\{y_w \mid w \in \mathcal{W}\}$ in $\Pi(\boldsymbol{p})$, we can obtain the following sufficient condition for $\Omega = 0$.

PROPOSITION 7 **(Optimality Condition)**. *Denote the optimal solution of ADP-1 by $\boldsymbol{p}^*$. If for any $w \in \mathcal{W}$, either $B_w \leq \sum_{v \in \mathcal{V}_w} \varphi e_{vw} (\mu_v - \hat{s}_v) p^*_{vw}$ or $B_w \geq \sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p^*_{vw}$ holds, then $\Omega = 0$.*

REMARK 4. Note that $\sum_{v \in \mathcal{V}_w} \varphi e_{vw} \mu_v p^*_{vw}$ is the expense of non-guaranteed campaign $w$ under plan $\boldsymbol{p}^*$ in the nominal case ($\tilde{s}_v = \mu_v$, $\forall v \in \mathcal{W}$), and $\sum_{v \in \mathcal{W}} \varphi e_{vw} (\mu_v - \hat{s}_v) p^*_{vw}$ is the expense of campaign $w$ under plan $\boldsymbol{p}^*$ in the worst case ($\tilde{s}_v = \mu_v - \hat{s}_v, \forall v \in \mathcal{W}$). Thus, the optimality condition essentially states that if under the optimal resource allocation plan, each campaign's budget is either not fully used in the nominal case or fully used even in the worse case, then the optimal solutions to ADP-1 and ADP-2 coincide. In general, we have the following performance guaranteed for the approximation.

PROPOSITION 8 **(Error Bound)**. *We have $\Omega \leq \sum_{v \in \mathcal{V}_\Gamma} \varphi \hat{s}_v \bar{e}_v$, where $\bar{e}_v = \max_{w \in \mathcal{W}_v} e_{vw}$, and $\mathcal{V}_\Gamma$ is a subset of $\mathcal{V}$ such that $|\mathcal{V}_\Gamma| = \lceil \Gamma \rceil$ and $\min_{v \in \mathcal{V}_\Gamma} \hat{s}_v \bar{e}_v \geq \max_{v \in \mathcal{V} \setminus V_\Gamma} \hat{s}_v \bar{e}_v$.*

REMARK 5. Considering the realistic situation of our problem, the error bound $\sum_{v\in\mathcal{V}_\Gamma} \varphi \hat{s}_v \bar{e}_v$ will not be large compared to the optimal objective function value of ADP-1. To illustrate, consider the case $|\mathcal{V}| = 10,000$ and $\varepsilon = 0.01$, and the corresponding $\Gamma = g^{-1}(|\mathcal{V}|, \varepsilon)$ is only nearly 234. Moreover, note that $\hat{s}_v$ is often considerably smaller than $\mu_v$ for segments of larger supply. Nevertheless, the error bound may perform badly if (1) the demand of guaranteed ad campaigns is very high (high sell-through) and the guaranteed ads and non-guaranteed ads compete intensively for the same resources (which means $\sum_{w\in\mathcal{W}_v} p_{vw}$ is much below 1; the proof of Proposition 8 shows that this will lead to an exaggeration of the error) and (2) supplies are too concentrated on a limited number of segments. We will comment on this further in the numerical studies. Henceforth, we focus on solving model ADP-2.

## 5. Clustering Algorithm

Although model ADP-2 is a linear program, solving it directly becomes computationally challenging when the number of viewer segments is rather large, which is often the case in practice. Therefore, for the purpose of real applications, a more efficient solution algorithm is required.

To overcome this challenge, we propose a clustering algorithm (Zipkin 1980, Turner 2012). This algorithm essentially partitions viewer segments into clusters, effectively reducing the problem scale; thus, the resulting problem is solved efficiently. In the remainder of this section, we first illustrate the clustering approach with an example and then present the clustered model, followed by a description of the steps involved in the clustering algorithm. The optimality gap of the algorithm is also characterized.

The notations introduced in Section 3.1 are extended and shown in Table 2. We illustrate the idea of clustering with an example. Suppose that there are two ad campaigns, $j$ and $w$, and five viewer segments, with campaign $j$ targeting segment $\{1,4,5\}$ and campaign $w$ targeting segments $\{2,3,5\}$. We partition the viewer segments into two clusters, $c_1$ and $c_2$, with viewer segments 1, 2, and 3 in cluster $c_1$ and the other viewer segments in cluster $c_2$. Therefore, the cluster set is $\mathcal{C} = \{c_1, c_2\}$, with $\mathcal{V}_{c_1} = \{1,2,3\}$, $\mathcal{V}_{c_2} = \{4,5\}$, $\mathcal{V}_{c_1 j} = \{1\}$, $\mathcal{V}_{c_1 w} = \{2,3\}$, $\mathcal{V}_{c_2 j} = \{4,5\}$, and $\mathcal{V}_{c_2 w} = \{5\}$. Furthermore, clusters $c_1$ and $c_2$ are both targeted by ad campaigns $j$ and $w$; thus, $\mathcal{C}_j = \mathcal{C}_w = \{c_1, c_2\}$. As shown in this example, viewer segments are partitioned into non-overlapping clusters; thus, there are typically considerably fewer clusters than viewer segments.

**Table 2** Extended notations

| Symbol | Description |
|---|---|
| $\mathcal{C}$ | set of clusters |
| $c(v)$ | the cluster containing viewer segment $v$ |
| $\mathcal{V}_c$ | set of viewer segments in cluster $c$ |
| $\mathcal{V}_{cj}$ | set of viewer segments in cluster $c$ and targeted by guaranteed campaign $j$ |
| $\mathcal{V}_{cw}$ | set of viewer segments in cluster $c$ and targeted by non-guaranteed campaign $w$ |
| $\mathcal{J}_c$ | set of guaranteed ad campaigns targeting cluster $c$ |
| $\mathcal{C}_j$ | set of clusters targeted by guaranteed ad campaign $j$ |
| $\mathcal{W}_c$ | set of non-guaranteed ad campaigns targeting cluster $c$ |
| $\mathcal{C}_w$ | set of clusters targeted by non-guaranteed ad campaign $w$ |

### 5.1. Clustered Model

For a given clustering $\mathcal{C}$, we first define the corresponding clustered model, which is denoted by ADP-3($\mathcal{C}$), and we then present some properties of the clustered model.

Let $\boldsymbol{q}(\mathcal{C}) = \{q_{cj} \mid j \in \mathcal{J}, c \in \mathcal{C}_j;\ q_{cw} \mid w \in \mathcal{W}, c \in \mathcal{C}_w\}$ be the clustered ad delivery plan, in which $q_{cj}$ and $q_{cw}$ are the proportions of impressions from cluster $c$ allocated to guaranteed ad campaign $j$ and non-guaranteed ad campaign $w$, respectively. Moreover, let $\boldsymbol{n}(\mathcal{C}) = \{n^d_{cj}, n^r_{cj} \mid j \in \mathcal{J}, c \in \mathcal{C}_j;\ n_{cw} \mid w \in \mathcal{W}, c \in \mathcal{C}_w\}$ be the set of clustered auxiliary variables, which correspond to auxiliary variables $\{m^d_{vj}, m^r_{vj}, m_{vw}\}$ in model ADP-2. The clustered model can be constructed from model ADP-2 in two steps.

In the first step, we confine $p_{vj}$ (resp. $p_{vw}$) to be equal for all viewer segments in the same cluster for any guaranteed ad campaign $j$ (resp. non-guaranteed ad campaign $w$), i.e.,

$$p_{vj} = q_{c(v)j}, v \in \mathcal{V}_j,\ \forall j \in \mathcal{J};\ \text{and}\ p_{vw} = q_{c(v)w},\ v \in \mathcal{V}_w,\ \forall w \in \mathcal{W}. \tag{24}$$

Similarly, auxiliary variables for all viewer segments in the same cluster are also confined to be equal, i.e.,

$$m^d_{vj} = \frac{n^d_{c(v)j}}{|\mathcal{V}_{c(v)j}|},\ m^r_{vj} = \frac{n^r_{c(v)j}}{|\mathcal{V}_{c(v)j}|}, v \in \mathcal{V}_j,\ \forall j \in \mathcal{J};\ \text{and}\ m_{vw} = \frac{n_{c(v)w}}{|\mathcal{V}_{c(v)w}|},\ v \in \mathcal{V}_w,\ \forall w \in \mathcal{W}. \tag{25}$$

This is clearly a restriction of the original problem and will lead to a sub-optimal solution. However, it will also reduce the problem scale.

Therefore, in the second step, we make a relaxation. The supply constraint after the restriction in step one is

$$\sum_{j \in \mathcal{J}_v} q_{c(v)j} + \sum_{w \in \mathcal{W}_v} q_{c(v)w} \leq 1,\ \forall v \in \mathcal{V}.$$

For each supply constraint on viewer segment $v$, we multiply it by $\mu_v$ on both sides and then sum them across viewer segments in each cluster $c \in \mathcal{C}$, leading to the following new supply constraint for model ADP-3$(\mathcal{C})$:

$$\sum_{j \in \mathcal{J}_c} \mu_{cj} q_{cj} + \sum_{w \in \mathcal{W}_c} \mu_{cw} q_{cw} \le \mu_c, \ \forall c \in \mathcal{C}, \text{ where } \mu_{cj} = \sum_{v \in \mathcal{V}_{cj}} \mu_v, \mu_{cw} = \sum_{v \in \mathcal{V}_{cw}} \mu_v, \ \mu_c = \sum_{v \in \mathcal{V}_c} \mu_v. \tag{26}$$

For all the other constraints with respect to each viewer segment $v \in \mathcal{V}$, we simply sum them across viewer segments in each cluster $c \in \mathcal{C}$.

Now, we present the formulation of ADP-3$(\mathcal{C})$ as follows:

$$\begin{aligned}
[\text{ADP-3}(\mathcal{C})] \ \max -\Gamma x + \sum_{w \in \mathcal{W}} h_w, & \\
\text{s.t. } |\mathcal{V}_c| x \ge \sum_{w \in \mathcal{W}_c} n_{cw}, \ \forall c \in \mathcal{C}, & \\
h_w - \sum_{c \in \mathcal{C}_w} n_{cw} \le \sum_{c \in \mathcal{C}_w} \varphi e_{cw} q_{cw}, \forall w \in \mathcal{W}, & \\
n_{cw} \le \varphi \hat{e}_{cw} q_{cw}, \ \forall c \in \mathcal{C}_w, w \in \mathcal{W}, & \\
h_w \le B_w, \ \forall w \in \mathcal{W}, & \\
\sum_{c \in \mathcal{C}_j} \varphi \mu_{cj} q_{cj} - \sum_{c \in \mathcal{C}_j} n^d_{cj} - \Gamma^d_j h^d_j \ge D_j, \ \forall j \in \mathcal{J}, & \\
|\mathcal{V}_{cj}| h^d_j + n^d_{cj} \ge \varphi \hat{s}_{cj} q_{cj}, \ \forall c \in \mathcal{C}_j, j \in \mathcal{J}, & \\
-\Gamma^r_j h^r_j + \sum_{c \in \mathcal{C}_j} n^r_{cj} \ge R_j, \ \forall j \in \mathcal{J}, & \\
-|\mathcal{V}_{cj}| h^r_j + n^r_{cj} \le -a^l_{cj}(q_{cj}) + b^l_{cj}(q_{cj}), \ \forall c \in \mathcal{C}_j, j \in \mathcal{J}, l \in \mathcal{L}, & \\
n^r_{cj} \le b^l_{cj}(q_{cj}), \ \forall c \in \mathcal{C}_j, j \in \mathcal{J}, l \in \mathcal{L}, & \\
\sum_{j \in \mathcal{J}_c} \mu_{cj} q_{cj} + \sum_{w \in \mathcal{W}_c} \mu_{cw} q_{cw} \le \mu_c, \ \forall c \in \mathcal{C}, & \\
0 \le q_{cj} \le 1, \ n^d_{cj} \ge 0, \ \forall c \in \mathcal{C}_j, j \in \mathcal{J}, & \\
0 \le q_{cw} \le 1, \ n_{cw} \ge 0, \ \forall c \in \mathcal{C}_w, w \in \mathcal{W}, & \\
h^d_j \ge 0, \ h^r_j \ge 0, \ \forall j \in \mathcal{J}, &
\end{aligned}$$

where $e_{cw} = \sum_{v \in \mathcal{V}_{cw}} e_{vw}(\mu_v - \hat{s}_v)$ and $\hat{e}_{cw} = \sum_{v \in \mathcal{V}_{cw}} e_{vw} \hat{s}_v$, $\forall c \in \mathcal{C}_w, w \in \mathcal{W}$; $\hat{s}_{cj} = \sum_{v \in \mathcal{V}_{cj}} \hat{s}_v$, $a^l_{cj}(q_{cj}) = \sum_{v \in \mathcal{V}_{cj}} a^l_v(q_{cj})$, and $b^l_{cj}(q_{cj}) = \sum_{v \in \mathcal{V}_{cj}} b^l_v(q_{cj})$, $\forall c \in \mathcal{C}_j, c \in \mathcal{J}$.

Note that ADP-3($\mathcal{C}$) remains as a linear program, but the problem scale depends on the number of clusters. Thus, the problem can be of much smaller scale compared to ADP-2 and allows better tractability.

Equations (24) and (25) map the solution to ADP-3($\mathcal{C}$) back to ADP-2. However, the relaxed constraints in ADP-3($\mathcal{C}$) cannot guarantee that the constraints in ADP-2 with respect to viewer segments are all satisfied in the mapped solution. If the mapped solution is infeasible for ADP-2, then we will need to refine the clustering, as is discussed in the next subsection.

### 5.2. Solution Procedure

**Clustering Initialization**. To start the clustering algorithm, we need a set of initial clusters, which is denoted by $\mathcal{C}^0$. Turner (2012) proposed to start with a mega-cluster, that is, all viewer segments initially belongs to one unique cluster. Conceivably, this initialization algorithm may not work well when the scale of instance is very large. For this reason, we propose an alternative algorithm that generates a specified number of initial clusters. The details are provided as the e-companion Section EC.7. However, as the limited test cases show in the e-companion Section EC.9, the performance of the alternative algorithm does not exhibit significant dominance over the mega-clustering algorithm. More effort is called for in future research.

**Clustering Refinement**. With an initial clustering $\mathcal{C}^0$, we solve ADP-3($\mathcal{C}^0$) and obtain a plan $\boldsymbol{q}(\mathcal{C}^0)$. However, as previously mentioned, the plan $\boldsymbol{p}(\mathcal{C}^0)$ mapped from $\boldsymbol{q}(\mathcal{C}^0)$ may be infeasible for ADP-2. Thus, we need to refine the clustering to eliminate such infeasibility. The entire refining procedure is protected by the following property.

PROPOSITION 9 **(Limited Infeasibility)**. *If a clustered plan $\boldsymbol{q}(\mathcal{C})$ is feasible for ADP-3($\mathcal{C}$), then the mapped solution $\boldsymbol{p}(\mathcal{C})$ according to equation (24) has the following property: for each cluster $c \in \mathcal{C}$, define $\mathcal{V}_c^s = \{v \in \mathcal{C} \mid$ supply constraint (20) is satisfied for viewer segment $v\}$, then, $\mathcal{V}_c^s \neq \emptyset$.*

The basic procedure for the clustering refinement is as follows. Denote the current clustering by $\mathcal{C}^1$. For the mapped plan $\boldsymbol{p}(\mathcal{C}^1)$, identify those clusters $c \in \mathcal{C}^1$ with $\mathcal{V}_c^s \neq \mathcal{V}_c$, and for each such cluster, split it into two clusters: $\mathcal{V}_c^s$ and $\mathcal{V}_c \setminus \mathcal{V}_c^s$. We cannot exclude the possibility (although we never encounter this in the experiments) that the mapped solution $\boldsymbol{p}(\mathcal{C})$ and $\boldsymbol{m}(\mathcal{C})$ is infeasible whereas for all clusters $\mathcal{V}_c^s = \mathcal{V}_c$, namely, all the supply constraints are satisfied. In this case, we will then randomly pick up some clusters to split.

Turner (2012) was able to show that if the optimal solution to a clustered model is also feasible for the original problem, then the solution is also optimal for the original problem. Simply speaking, this is because the duals to both the clustered and original problems have no constraints other than simple bounds of dual variables; therefore, the optimal solution to the dual of the clustered problem can be transformed into a feasible solution to the dual of the original problem while keeping the same objective value. Unfortunately, our clustering algorithm does not have this nice property — our original problem (ADP-2) has many auxiliary variables introduced for transforming chance constraints, which correspond to many constraints in the dual problem. Nevertheless, the clustering refinement is guaranteed to terminate in a finite number of steps, according to the following simple, yet important, proposition.

PROPOSITION 10 **(Clustering Convergence)**. *The clustering refinement procedure will converge to a feasible solution in a finite number of steps if the original problem ADP-2 is feasible.*

**Heuristic Method**. Although Proposition 10 guarantees the convergence of the clustering refinement, for very large problem scales, the convergence may be time consuming. Therefore, we allow the refinement to stop earlier when the supply over-allocation is not serious.

Suppose that we stop with an infeasible plan $\boldsymbol{p}'$. We attempt to convert $\boldsymbol{p}'$ into a feasible plan using a two-phase heuristic method. The first phase iteratively searches for a feasible plan tailored for guaranteed campaigns (all the proportions of supply allocated to non-guaranteed campaigns are manually set as zero), and each iteration consists of two steps, with the first step addressing the issue of supply over-allocation and the second step addressing the newly introduced violations of reach and delivery constraints. Successfully finding a feasible plan in the first phase will invoke the second phase; otherwise, the heuristic is terminated. In the second phase, we re-compute the proportions of supply allocated to the non-guaranteed campaigns in the feasible plan by solving an auxiliary linear program. The details of the heuristic method are provided in the e-companion Section EC.9.

### 5.3. Upper Bound

For very large-scale cases, due to time limitation, we may want to stop the clustering algorithm earlier if we have found a very good solution. To measure the optimality gap, considering that it is difficult to establish the optimal value of ADP-2, we seek to find its upper bound. For this purpose, we present the following linear program, which is denoted by RADP-2.

$$[\text{RADP-2}] \quad \max \sum_{w \in \mathcal{W}} u_w,$$

$$
\begin{aligned}
\text{s.t. } & u_w \leq \sum_{v \in \mathcal{V}_w} \varphi e_{vw} \left(\mu_v - \hat{s}_v \bar{\phi}_v\right) p_{vw}, \ \forall w \in \mathcal{W}, \\
& u_w \leq B_w, \ \forall w \in \mathcal{W}, \\
& \sum_{v \in \mathcal{V}_j} \varphi \left(\mu_v - \hat{s}_v \bar{\theta}_{vj}\right) p_{vj} \geq D_j, \ \forall j \in \mathcal{J}, \\
& \sum_{v \in \mathcal{V}_j} u_{vj} \geq R_j, \ \forall j \in \mathcal{J}, \\
& u_{vj} \leq -\bar{\eta}_{vj} a_v^l(p_{vj}) + b_v^l(p_{vj}), \ \forall v \in \mathcal{V}_j, j \in \mathcal{J}, l \in \mathcal{L}, \\
& \sum_{j \in \mathcal{J}_v} p_{vj} + \sum_{w \in \mathcal{W}_v} p_{vw} \leq 1, \ \forall v \in \mathcal{V}, \\
& 0 \leq p_{vj} \leq 1, \ \forall v \in \mathcal{V}, \ j \in \mathcal{J}_v; \ 0 \leq p_{vw} \leq 1, \ \forall v \in \mathcal{V}, \ w \in \mathcal{W}_v,
\end{aligned}
$$

where

$$
\bar{\phi}_v = \begin{cases} 1 & \text{if } v \in \mathcal{V}(\Gamma) \\ \Gamma - \lfloor \Gamma \rfloor & \text{if } v = v^*(\Gamma) \\ 0 & \text{o.w.} \end{cases}, \ \bar{\theta}_{vj} = \begin{cases} 1 & \text{if } v \in \mathcal{V}_j(\Gamma_j^d) \\ \Gamma_j^d - \lfloor \Gamma_j^d \rfloor & \text{if } v = v^*(\Gamma_j^d) \\ 0 & \text{o.w.} \end{cases}, \ \bar{\eta}_{vj} = \begin{cases} 1 & \text{if } v \in \mathcal{V}_j(\Gamma_j^r) \\ \Gamma_j^r - \lfloor \Gamma_j^r \rfloor & \text{if } v = v^*(\Gamma_j^r) \\ 0 & \text{o.w.} \end{cases},
$$

and in the above, $\mathcal{V}(\Gamma) \subseteq \mathcal{V}$ is a set with $\lfloor \Gamma \rfloor$ viewer segments with the largest $\hat{s}_v$, i.e., for any $v \in \mathcal{V}(\Gamma)$, $\hat{s}_v \geq \max_{v' \in \mathcal{V} \setminus \mathcal{V}(\Gamma)} \hat{s}_{v'}$, $v^*(\Gamma) = \arg\max_{v \in \mathcal{V} \setminus \mathcal{V}(\Gamma)} \hat{s}_v$; similarly, for each $j \in \mathcal{J}$, $\mathcal{V}_j(\Gamma_j^d) \subseteq \mathcal{V}_j$ is a set with $\lfloor \Gamma_j^d \rfloor$ elements with the largest $\hat{s}_v$, $v^*(\Gamma_j^d) = \arg\max_{v \in \mathcal{V}_j \setminus \mathcal{V}_j(\Gamma_j^d)} \hat{s}_v$, and the definitions of $\mathcal{V}_j(\Gamma_j^r)$ and $v^*(\Gamma_j^r)$ are similar with that of $\mathcal{V}_j(\Gamma_j^d)$ and $v^*(\Gamma_j^d)$.

PROPOSITION 11 **(Clustering Bound)**. *Denote the optimal value of RADP-2 by $Z^{RADP-2}$. We have $Z^{ADP-2} \leq Z^{ADP-1} \leq Z^{RADP-2}$.*

We thus have an upper bound for the optimal value of ADP-2. Throughout the computation, we may stop earlier if we have found a very satisfactory solution that is close to the upper bound. Denote a lower bound to $Z^{ADP-2}$ by $\underline{Z}$, Proposition 11 also provides an upper bound for $\Omega$: $\Omega \leq Z^{RADP-2} - \underline{Z}$.

## 6. Ad Serving Implementation

By solving the ad planning model, we have obtained an ad delivery plan. In this section, we detail the ad serving procedure, i.e., how to implement the ad delivery plan in real time. We precede our discussion with a description of the data sets to be used for the numerical studies, including two real data sets that motivate the following discussions.

### 6.1. Data Sets and Setting Reach

For the numerical studies, we will use three data sets, including one randomly generated data set and two real data sets, with a total of 14 test cases. Among them, A1∼A6 are randomly generated cases, produced by a method adapted from Turner (2012) as detailed in the e-companion Section EC.10. Note that these 6 test cases are characterized by sell-through and targeting intensity. Sell-through is the ratio of the total demand from all guaranteed ad campaigns to the total expected impressions of viewer segments targeted by guaranteed ad campaigns. Targeting intensity is the average percentage of viewer segments targeted by an ad campaign. Intuitively, a higher sell-through implies a greater difficulty in achieving feasibility, and a higher targeting intensity means a greater problem complexity. The random test cases are chosen to represent different combinations of the defining characteristics.

**Table 3** Test data

| Test case | Num. viewer segments | Num. guaranteed campaigns | Num. non-guaranteed campaigns | Sell-through (%) | Targeting intensity (%) |
|---|---|---|---|---|---|
| A1 | 10,000 | 100 | 100 | 31.8 | 18.1 |
| A2 | 10,000 | 100 | 100 | 55.1 | 18.4 |
| A3 | 10,000 | 100 | 100 | 80.2 | 3.4 |
| A4 | 100,000 | 100 | 100 | 31.7 | 17.6 |
| A5 | 100,000 | 100 | 100 | 54.2 | 18.3 |
| A6 | 100,000 | 100 | 100 | 79.2 | 3.3 |
| B1 | 13,133 | 232 | 214 | 47.4 | 1.5 |
| B2 | 13,343 | 222 | 224 | 33.5 | 1.6 |
| B3 | 11,682 | 207 | 212 | 38.3 | 1.7 |
| B4 | 13,526 | 227 | 228 | 38.8 | 1.6 |
| C1 | 35,563 | 322 | 83 | 72.6 | 18.8 |
| C2 | 31,184 | 347 | 101 | 59.7 | 24.5 |
| C3 | 30,862 | 312 | 124 | 56.4 | 23.4 |
| C4 | 29,190 | 490 | 118 | 64.4 | 20.3 |

We also use two real data sets, one (test cases B1∼B4) from an SNS firm and the other (test cases C1∼C4) from a video advertising firm, each of which the first author spent four months working within. The latter, FastAdv (whose real name has been withheld for confidentiality), is a major digital video advertising solution provider, and its ad serving platform is used by many TV networks and major video distributors to serve ads alongside their content online and on mobile.

Headquartered in the U.S., its operations and engineering teams are located in Asia. Around 300 engineers are currently working for supply/demand forecasting, ads planning and serving, system maintenance, and product design and test. The company has helped a large number of broadcasters manage the delivery of ads for live, news-breaking events, and programming, such as sport events and movie streaming.

In the data set used here, we focus on an EU-based TV network, STV (the actual name is disguised), one of FastAdv's clients. STV streams its content online and sells its advertising inventory to big brands and other advertisers. It uses FastAdv for targeted advertising through dynamic insertion of ads into its TV programs. To put simply, STV outsources advertisement delivery to FastAdv. Advertisers negotiate contracts with STV, which specify all the requirements in terms of the number of impressions and target segments among others. Then STV hands over FastAdv to fulfill these contracts by displaying the ads over time and in the meantime fulfilling all the requirements that are similar to those of our earlier model.

**Setting Reach**. The FastAdv data set is more sophisticated than the SNS data set. Specifically, there are three situations regarding reach specification in FastAdv. (1) Some upfront advertisers specify their reach requests. In this case, we then have a natural *hard* reach constraint. (2) Some advertisers do not fill in their reach requests, but they specify frequency caps for their ad campaigns. A frequency cap specifies the maximum number of times that an ad campaign can be shown to the same viewer over a period. Note that all guaranteed ads specify the number of impressions. Combining the impressions and frequency caps, we can have a natural lower bound for a campaign's reach:

$$\text{minimum reach} = \text{number of impressions / frequency cap.}$$

Such a lower bound is also a hard constraint for fulfilling advertisers' frequency cap specification. (3) However, the rest of the advertisers specify neither reach nor frequency caps. The firm then uses a very loose frequency cap (specifically, the maximum frequency cap found from other campaigns $\bar{r} + 1$) to compute a target reach. The purpose is to prevent from delivering extremely undesirable reach outcomes, which would shock the advertisers when they see the delivery reports. Note that a very loose frequency cap has little impact on the objective of maximizing spot market revenue.

The SNS provider is an NYSE listed company that operates a social networking service and internet finance business in China. Its SNS website and mobile application had approximately 240 million activated users as of December 31, 2016. At the time of our data collection, the firm had over 50 million active SNS users per month and derived its major revenue from online display advertising. As we have described in the introduction, the firm attempts to maximize its spot-market revenue while also fulfilling its commitment with guaranteed ads. To grow its business in the upfront market, the firm hopes to build up its reputation of quality advertising, and for this reason, it aims to have a reasonable reach for each guaranteed campaign given that this does not sacrifice its spot-market revenue considerably. Therefore, the firm needs to try different reaches and then select one that is acceptable to itself. Nevertheless, such a reach target is considered a *soft* constraint. Motivated by the practice of FastAdv, we generate the SNS firm's initial reach inputs by imposing a loose frequency cap (specifically, we set the frequency cap as the ratio of total supply to the total number of unique individuals for test cases B1$\sim$B4, and also for A1$\sim$A6) on ad campaigns, and then we gradually increase the reach to generate an efficiency frontier between reach and revenue, which allows the firm to select its preferred reach level.

### 6.2. Ad Serving Method

To conform to Assumption 1, we need to maintain the randomness of ad serving. Therefore, we propose the following *slate* method for ad serving.

For each ad viewer segment $v$, we first generate a slate that has $\Psi$ slots. We then assign $\Psi \cdot p_{vj}$ slots of the slate to each guaranteed ad campaign $j \in \mathcal{J}_v$. If $\Psi \cdot p_{vj}$ is not an integer, then we assign $\lfloor \Psi \cdot p_{vj} \rfloor$ slots to ad campaign $j$ and an additional one with probability $(\Psi \cdot p_{vj} - \lfloor \Psi \cdot p_{vj} \rfloor)$. In the same way, we assign slots of the slate to each non-guaranteed ad campaign $w \in \mathcal{W}_v$. The slot sequence is shuffled after all campaigns are assigned. The slate length $\Psi$ is chosen such that it is long enough to ensure each ad campaign receives a sufficient number of slots but not too long to affect running efficiency. In the experiment, we set $\Psi = \lceil \mu_v/3 \rceil$ if $\mu_v < 3000$ and $\Psi = 1000$ otherwise. Ad serving is then implemented based on the slate. Specifically, the impressions are assigned one by one in the order of the slot sequence. The slate is regenerated every $\Psi$ impressions following the same procedure to further ensure randomness.

Any model is just an approximation to reality, and ours is no exception. Aside from frequency caps, exclusivity constraint is also not accounted for by our planning model, yet both exist in the FastAdv data set. With exclusivity constraints, advertisers specify that their ad campaigns cannot be shown together with some other ad campaigns (usually from competing brands) in the same video. Nevertheless, the above slate procedure can be adapted to incorporate the two constraints.

Briefly speaking, when an ad display opportunity arises but the scheduled ad campaign cannot be displayed due to violation of either frequency cap constraint or exclusivity constraint, we first try to search for the next feasible ad campaign on the slate. If one is found, we then swap the current ad campaign with the feasible ad campaign. If none can be found, we then generate a new slate and append it to the current slate (thus, we maintain no more than three slates at any time) and continue the above search procedure; we perform this step because feasible ad campaigns may likely be in the executed part of the current slate. If still no feasible ad campaign can be found, we then give up and choose to display either a compatible campaign (satisfying viewer targeting, frequency cap and exclusivity constraints, if any) or some charity (free) ads.

With the above slate method, the ad serving procedure is easy and efficient to implement, while maintaining randomness and ensuring that the percentages of impressions allocated are basically the same as stipulated in the plan. Finally, we can make a simple yet important adjustment during implementation. Whenever we find a guaranteed ad campaign has reached its impressions target and reach target or a non-guaranteed ad campaign has used up its budget, we will remove the ad campaign from the slate and assign the spare ad resource to the remaining ad campaigns proportionally. As we will see shortly, the adjustment is effective because such situations often arise due to the conservatism of the robust model.

### 6.3. Experiment Setup

To set up the ad planning model ADP-0, some parameters of the model deserve our special attention.

**Segment Supplies Specification**. For test cases A1$\sim$A6, we first randomly generate the mean segment supply $\mu_v$ following the method in Turner (2012), as described in the e-companion Section EC.10, and then let the half interval width $\hat{s}_v = \zeta\mu_v$, where $\zeta \sim$ Uniform(5%, 30%). For test cases B1$\sim$B4 from the SNS firm and C1$\sim$C4 from the video advertising firm, since we know the true supply (denoted by $s_v$), we first set the half interval width $\hat{s}_v = \zeta s_v$, where $\zeta \sim$ Uniform(5%, 30%), and then randomly generate the mean supply by letting $\mu_v \sim \text{Uniform}(s_v - \hat{s}_v, s_v + \hat{s}_v)$. Note that by this setting, the true supply $s_v$ must lie in the interval $[\mu_v - \hat{s}_v, \mu_v + \hat{s}_v]$. The underlying assumption is that the firm can make a reasonably good prediction on segment supplies, which is the case as of today. In the real practice of the video advertising company, over 100 engineers are working on forecasting. They are able to make quite satisfactory predictions with sophisticated machine learning techniques. In Section EC.11 of the e-companion, we vary the half interval width $\hat{s}_v$ to see the effect of prediction accuracy on the solution performance. As the prediction accuracy increases, the model becomes less conservative (as reflected by the decreasing difference between the percentage of unfilled campaigns and the specified $\varepsilon_j^d$ in the model), and the revenue from non-guaranteed ads increases.

**Specifying Risk Levels**. For our DRCC model, we need to specify risk levels for the revenue constraint, delivery constraints, and reach constraints, i.e., $\varepsilon, \varepsilon_j^d, \varepsilon_j^r$ in model ADP-0. We use the following guidelines: for hard constraints such as delivery constraints and also the reach constraints specified by advertisers or derived from advertisers' frequency caps, we set a lower risk level, and let $\varepsilon_j^d = 5\%$, $\varepsilon_j^r = 5\%$; for soft constraints, namely, reach constraints set by the publisher, we set a higher risk level, and let $\varepsilon_j^r = 10\%$. Moreover, for a few campaigns with impressions target below 1000, we set $\varepsilon_j^d = 1\%$, $\varepsilon_j^r = 1\%$ because these campaigns receive a very small proportion of supply and are susceptible to ad serving uncertainty. We simply choose $\varepsilon = 10\%$ for no particular reason, as this depends on the publisher's risk appetite.

### 6.4. Experimental Results

We now describe the experiments based on the 14 test cases. Our purposes are to assess the effectiveness of the ad delivery plan and the ad serving procedure, learn the conservatism of the approximations, explore the trade-off between upfront market and spot market, and examine the performance of our solution algorithms.

All experiments were implemented in Java, and we used MOSEK (version 9), an off-the-shelf solver for convex optimization models, to solve the models. The experiment was conducted on a PC equipped with an Intel i7 processor (clocked at 3.2 GHz) and 32GB RAM, running 64-bit Windows 10 operating system. Notably, for 8 out of the 14 test cases, i.e., A1$\sim$A3, A6, and B1$\sim$B4, we are able to directly solve model ADP-2 to optimality. Note that although test case A6 is of large scale, it has a lower targeting intensity and is thus easier to solve. For the remaining larger test cases, we need to resort to the clustering algorithm and solve model ADP-3($\mathcal{C}$).

Note that we do not have the real user arrival information for the first 10 test cases (A1$\sim$B4). To evaluate the performance of the obtained ad delivery plan for those 10 cases, we simulate real-time ad serving based on our plan for 1000 runs. In each run, for each viewer segment, we first randomly generate the impression supply (the number of user arrivals) according to its distribution. Then, based on Assumption 1 (all users in a segment are homogeneous), we draw each user arrival uniformly from all the users of the segment. The ad serving is conducted based on our slate ad serving method described earlier, and after the ad serving is completed, we record the total revenue gained from all non-guaranteed ad campaigns and the fill rate of the impression demand for each guaranteed ad campaign. For FastAdv test cases, since we have the detailed historical information, we directly implement our plan on the real data.

**Robustness of the Plan**. Table 4 reports the robustness of our plan in fulfilling the guaranteed delivery for ad campaigns in the upfront market. Note that the results for test cases A1$\sim$B4 are the average over 1000 simulations, while for test cases C1$\sim$C4 we run the simulation only once by replicating the history. Recall that we set $\varepsilon_j^d = 5\%$ for most campaigns and $\varepsilon_j^d = 1\%$ for smaller campaigns with impressions target below 1000. First, we can see that the generated plan is very robust and for almost all test cases, the fulfillment rate (here we measure the fulfillment only by impression delivery because reach is satisfied 100% due to the loose specification) is over 95%. Second, for those unfulfilled campaigns, the extent of under-delivery is rather low, mostly much below 5%. Third, for test cases A3 and B1$\sim$B4 because of their low targeting intensity, we have set for many of their campaigns $\varepsilon_j^d = 1\%$, and therefore, their fulfillment rates are all 100%. Overall, the generated plan has delivered its promise. Note that the reported results follow the slate ad serving procedure, without adjusting the delivery plan in the process to release and reallocate the ad resource allocated to the fulfilled guaranteed or non-guaranteed ads, which will distort the comparison.

**Table 4** Robustness of ad delivery

| Test case | Unfulfilled campaigns (%) | Unfulfilled demand (%) | Test case | Unfulfilled campaigns (%) | Unfulfilled demand (%) |
|---|---|---|---|---|---|
| A1 | 0.21 | 0.17 | B2 | 0 | 0 |
| A2 | 2.49 | 0.16 | B3 | 0 | 0 |
| A3 | 0 | 0 | B4 | 0 | 0 |
| A4 | 0.05 | 0.06 | C1 | 3.42 | 1.14 |
| A5 | 1.88 | 0.07 | C2 | 2.88 | 1.76 |
| A6 | 0.02 | 0.09 | C3 | 2.24 | 1.07 |
| B1 | 0 | 0 | C4 | 3.88 | 3.29 |

Another indication from Table 4 is that our approximations of the chance constraint seem very tight, especially for the FastAdv test cases. In using the chance-constrained program, setting risk levels can be critical, especially when safe approximations are involved. If the approximations are very conservative, then we will need to set the risk levels much higher than the standard levels 5% or 10%. To assess the conservatism of our method, we take case C4 as an example. The resulting percentages of unfulfilled campaigns are 9.80%, 7.76%, and 3.88% respectively when $\varepsilon_j^d = 15\%, 10\%, 5\%$. From this example, we can see that our chance constraint approximation can be considered reasonably tight. When setting risk levels, a slight amplification of the intended risk level will be fine, or one may even go straight with no amplification. (As campaigns run multiple days, slight under-delivery in one period can be compensated in the next day by adjusting that day's target upward. Hence, under-delivery in one day will not necessarily lead to overall non-fulfillment.)

**Frequency Cap and Exclusivity**. FastAdv data set contains constraints that are not modeled directly in our ad planning model ADP-0. A simple incorporation of unmodeled constraints, especially the exclusivity constraints, will no doubt lead to some infeasibility. However, our robust approach actually allows certain redundancy that can be used to deal with such constraints. Table 5 reports the results of implementing our plan to incorporate the constraints on test cases C1∼C4. There are three scenarios according to whether we incorporate the constraints and whether we adjust the ad serving to reallocate ad resources originally dedicated to fulfilled ad campaigns; the scenario with exclusivity constraints but without adjustment is not reported because the infeasibility is very high (over 10%). The first scenario corresponds to the original optimization model. In the second scenario, adjustment is made to better utilize the resources. Consequently, we see that all three performance measures (unfulfilled campaigns, unfulfilled demand, and realized revenue) are improved. In the third scenario, while making adjustment, we incorporate frequency and exclusivity, which inevitably make the performance related to guaranteed ads worse compared with the second scenario because essentially the ad resources become less effective. However, this leaves more ad resources to the spot market, and the revenue from non-guaranteed ads improves. In other words, scenarios 2 and 3 both make adjustment. Scenario 2 dominates scenario 1, while scenario 3 performs the best in terms of spot-market revenue, but not as good in terms of the performance of guaranteed ads.

It is noteworthy that for the four real test cases, after incorporating all the constraints in ad serving and with adjustment, our approach recorded on average about 10% more revenue than the delivery approach of the firm.

**Trade-off Between Revenue and Reach**. If the advertisers specify their targeted reach, then we consider the reach constraints as compulsory, hard constraints. Otherwise, the publisher may

**Table 5** Dealing with frequency cap and exclusivity

| Test case | Scenario | Unfulfilled campaigns (%) | Unfulfilled demand (%) | Realized revenue | Guaranteed revenue ($t^*$) |
|---|---|---|---|---|---|
| C1 | 1: w/o constraints, w/o adjustment | 3.42 | 1.14 | 48676 | 46441 |
| | 2: w/o constraints, w/ adjustment | 1.55 | 4.93 | 52255 | |
| | 3: w/ constraints, w/ adjustment | 1.86 | 23.16 | 52957 | |
| C2 | 1: w/o constraints, w/o adjustment | 2.88 | 1.76 | 49864 | 47714 |
| | 2: w/o constraints, w/ adjustment | 0.00 | 0.00 | 53331 | |
| | 3: w/ constraints, w/ adjustment | 3.75 | 8.51 | 53711 | |
| C3 | 1:w/o constraints, w/o adjustment | 2.24 | 1.07 | 37500 | 36218 |
| | 2: w/o constraints, w/ adjustment | 0.00 | 0.00 | 39176 | |
| | 3: w/ constraints, w/ adjustment | 0.96 | 11.17 | 39699 | |
| C4 | 1: w/o constraints, w/o adjustment | 3.88 | 3.29 | 30907 | 30085 |
| | 2: w/o constraints, w/ adjustment | 0.61 | 3.59 | 31713 | |
| | 3: w/ constraints, w/ adjustment | 3.27 | 10.71 | 31831 | |

want to voluntarily impose reach constraints for the purpose of ensuring advertising quality. However, the publisher does not expect such voluntary action to jeopardize its spot-market profitability. Therefore, a reach-revenue efficient frontier will help the publisher choose the ideal reach level. For this purpose, we first set an initial loose-reach target for each campaign and then gradually increase it. We solve a few such resulting models to construct the efficient frontier. The initial reach is set following the way described in Section 6.1. For each test case, we first solve model ADP-2 with the initial reach to obtain an ad delivery plan, and run 1000 simulations of ad serving based on the obtained plan. Next, we increase the reach for all the campaigns gradually and repeat the above steps iteratively.

Figure 1 shows the variations of the average total realized revenue from the spot market under different reach commitment values on test case A3; the other test cases produce similar results. The trade-off between revenue and reach is explicitly captured in the figure. As expected, increasing reach results in decreased revenue. The publisher then may pick up a point from the efficient frontier. Among other possibilities (such as boosting reputation in a short period of time), a target reach may be chosen such that an even larger reach will lead to a sharp decrease in revenue.

**Clustering Algorithm**. We now examine the efficiency and effectiveness of the clustering algorithm. Table 6 reports the performance of the clustering algorithm on the six test cases that model

**Figure 1** Revenue and reach in test case A3

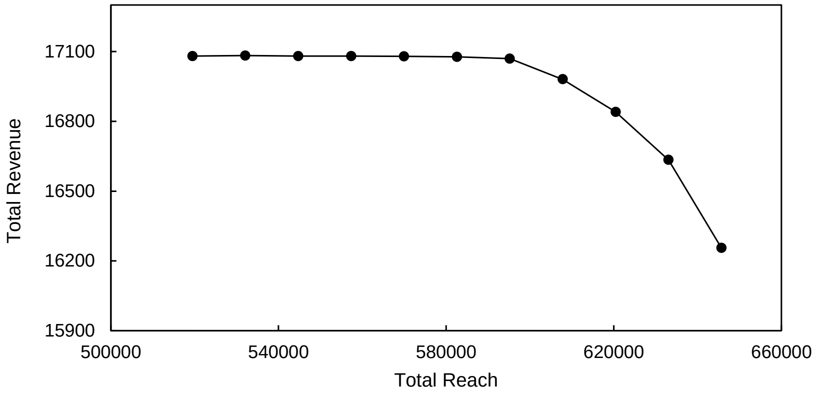

**Table 6** Performance of clustering algorithm

| Test case | Obj value | UB_RADP | Gap (%) |
|---|---|---|---|
| A4 | 657976 | 687370 | 4.47 |
| A5 | 452574 | 489008 | 8.05 |
| C1 | 48007 | 49461 | 3.03 |
| C2 | 49335 | 50934 | 3.24 |
| C3 | 37071 | 38200 | 3.05 |
| C4 | 30382 | 31211 | 2.73 |

ADP-2 cannot be solved directly. The optimality gap to the upper bound given by model RADP-2 ranges from 2.73% to 8.05%, which is reasonable, considering that RADP-2 is an upper bound. To make more sense of this, on the other 8 test cases of which we know the optimal solutions, RADP-2 gives upper bounds on average 2.28% higher than the optimal objective values, ranging from 0.61% to 7.19%.

**Error Bound**. Finally, we comment on the performance loss from ADP-1 to ADP-2. Proposition 8 gives an upper bound on the absolute performance loss. The ratio of the upper bound to the objective value of our solution gives an upper bound on the relative error because our objective value is a lower bound to $Z^{ADP-1}$. We report this relative error for all test cases in the column "EB" in Table 7. We note that the error bounds vary significantly, from less than 1% to nearly 15%. However, as noted in the earlier remark, the error bound by Proposition 8 can be an exaggeration of the true loss when the sell-through is high or the supplies are over-concentrated, which is exactly the case of test cases with high error bound, e.g., B1, C1, and C2. From a different angle, according to Proposition 11, the performance loss from ADP-1 to ADP-2 is also bounded by $(Z^{RADP-2} - \underline{Z})$, where $\underline{Z}$ here can be the objective value of our solution. Therefore, $(Z^{RADP-2} - \underline{Z})/\underline{Z}$ also gives an upper bound for the performance loss, and we report it in the column "EB-1" in Table 7. The minimum of the two columns "EB" and "EB-1" gives a tighter measure of the performance from ADP-1 to ADP-2, which is quite reasonable. Therefore, we are basically safeguarded to focus on solving ADP-2.

**Table 7** Error bound

| Test case | EB (%) | EB-1 (%) | min{EB, EB-1} | Test case | EB (%) | EB-1 (%) | min{EB, EB-1} |
|---|---|---|---|---|---|---|---|
| A1 | 1.39 | 1.10 | 1.10 | B2 | 4.82 | 1.38 | 1.38 |
| A2 | 1.88 | 1.86 | 1.86 | B3 | 5.36 | 1.77 | 1.77 |
| A3 | 3.65 | 7.19 | 3.65 | B4 | 5.91 | 1.89 | 1.89 |
| A4 | 0.51 | 4.47 | 0.51 | C1 | 8.09 | 3.03 | 3.03 |
| A5 | 0.77 | 8.05 | 0.77 | C2 | 15.10 | 3.24 | 3.24 |
| A6 | 1.31 | 0.61 | 0.61 | C3 | 4.62 | 3.05 | 3.05 |
| B1 | 7.25 | 2.48 | 2.48 | C4 | 3.51 | 2.73 | 2.73 |

## 7. Conclusions

Online display advertising publishers face the problem of planning their ad delivery, which is critical to both their short-term and long-term profitability. In particular, planning for advertising resources needs to consider several objectives. First, the publisher is required to fulfill its contractual promises to the advertisers in the upfront market. Second, it is of great importance for the publisher to maximize the revenue from the spot market, because the revenue in the upfront market is already fixed. Third, to maintain a long-term relationship with advertisers in the upfront market, the publisher needs to ensure that the advertising is effective, as measured by the audience reach.

The current way of selling ad resources suffers from drawbacks. For one thing, auction as a way of selling ad resources at the spot market has been recognized to be inefficient due to factors such as the advertisers' daily budgets (Parkes and Sandholm 2005, Benisch et al. 2009). For websites such as social network sites or others with relatively predictable traffic, a planning approach promises considerable potential. Second, ad resource allocation typically lacks coordination between the upfront market and spot market, although the same ad resources are shared by both markets.

Therefore, in this study, we proposed to integrate ad delivery planning of the upfront market and spot market. We formulated the problem as a DRCC program and designed solution methods. In the model, the guaranteed delivery and reach in the upfront market are modeled as constraints, and the spot market revenue is maximized. Representativeness, although not directly incorporated into the model, is reflected to a certain extent; to achieve robustness and a high reach, a good plan tends to allocate the supply of each viewer segment to multiple ad campaigns and serve each ad campaign by many viewer segments, and hence, representativeness is also implicitly favored by our model. The extensive numerical studies demonstrate the effectiveness of the approach.

Note that a prerequisite of using the model in practice is that the impression supplies are relatively predictable, and this is the case in most situations of display advertising such as social network sites. Further improved forecasting capability will no doubt benefit the application of our approach. Moreover, implementing our approach on an even larger scale (consider Facebook, for example) may require publishers to further decompose their problems into subproblems of a smaller scale, e.g., from continent level to country level.

## Acknowledgments

The authors are thankful to Prof Chung-Piaw Teo, the Associate Editor and two anonymous referees for their constructive comments. They also thank Zhengfa Dang, Jack Deng, Wei Li, and Haozhi Liu for sharing their industry experience and enabling the test of the proposed model and method. Huaxiao Shen is supported by National Natural Science Foundation of China (Project No. 71801231) and Humanities and Social Science Fund of Ministry of Education of China (Project No. 17YJC630120). Yanzhi Li acknowledges the support of Strategic Research Grant (Project No. 7004269 and 7004785) from City University of Hong Kong. Youhua Chen is supported by National Natural Science Foundation of China (Project No. 71390335). Kai Pan is supported by The Research Grants Council of Hong Kong (Project No. 15501319).

## Author Biographies

**Huaxiao Shen** received his PhD from City University of Hong Kong. His research interests include business analytics and interface research between operations and marketing.

**Yanzhi Li** is a Professor of Management Sciences and Marketing at City University of Hong Kong. He received a bachelor's degree from Tsinghua University and his PhD from Hong Kong University of Science and Technology. Yanzhi Li's research focuses on applying analytics to solve challenging problems arising from operations, logistics, healthcare, revenue management and marketing.

**Youhua (Frank) Chen** is currently Chair Professor and Dean of College of Business, City University of Hong Kong. He received his PhD from the University of Toronto, Canada. His current research projects cover healthcare management, data analytics in business operations, and applications of machine learning in supply chain management. He is also coordinating a team of registered nurses and social workers to innovate the community-based elderly care.

**Kai Pan** is an Assistant Professor in the Department of Logistics and Maritime Studies of the Faculty of Business at the Hong Kong Polytechnic University. He received a bachelor's degree from Zhejiang University and his PhD from the University of Florida. His research interests include stochastic and discrete optimization, robust and data-driven optimization, dynamic programming, and their applications in energy market, supply chain, marketing, and transportation.

## References

Abhishek V, Hosanagar K (2013) Optimal bidding in multi-item multislot sponsored search auctions. *Oper. Res.* 61(4):855–873.

Araman VF, Popescu I (2010) Media revenue management with audience uncertainty: Balancing upfront and spot market sales. *Manufacturing Service Oper. Management* 12(2):190–212.

Ardestani-Jaafari A, Delage E (2016) Robust optimization of sums of piecewise linear functions with application to inventory problems. *Oper. Res.* 64(2):474–494.

Balseiro SR, Feldman J, Mirrokni V, Muthukrishnan S (2014) Yield optimization of display advertising with ad exchange. *Management Sci.* 60(12):2886–2907.

Ben-Tal A, Den Hertog D, Vial JP (2015) Deriving robust counterparts of nonlinear uncertain inequalities. *Math. Programming* 149(1-2):265–299.

Benisch M, Sadeh NM, Sandholm T (2009) Methodology for designing reasonably expressive mechanisms with application to ad auctions. *Proc. 21th Internat. Joint Conf. Artificial Intelligence*, 46–52 (Morgan Kaufmann Publishers, San Francisco, CA).

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.

Bhalgat A, Feldman J, Mirrokni V (2012) Online allocation of display ads with smooth delivery. *Proc. 18th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining*, 1213–1221 (ACM, New York).

Bharadwaj V, Chen P, Ma W, Nagarajan C, Tomlin J, Vassilvitskii S, Vee E, Yang J (2012) SHALE: an efficient algorithm for allocation of guaranteed display advertising. *Proc. 18th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining*, 1195–1203 (ACM, New York).

Charnes A, Cooper WW (1959) Chance-constrained programming. *Management Sci.* 6(1):73–79.

Chen J, Liu D, Whinston AB (2009) Auctioning keywords in online search. *J. Marketing* 73:125–141.

Chen W, Sim M, Sun J, Teo CP (2010) From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Oper. Res.* 58(2):470–485.

Chen X, Sim M, Sun P (2007) A robust optimization perspective on stochastic programming. *Oper. Res.* 55(6):1058–1071.

Chen YJ (2017) Optimal dynamic auctions for display advertising. *Oper. Res.* 65(4):897–913.

Deza A, Huang K, Metel MR (2015) Chance constrained optimization for targeted internet advertising. *Omega* 53:90–96.

Edelman B, Ostrovsky M, Schwarz M (2007) Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *Am. Econ. Rev.* 97(1):242–259.

eMarketer (2019) Display ad spending, worldwide. Accessed August, 2019, https://forecasts-na1.emarketer.com/5a4e7234d8690c0c28d1f2e4/5a4e66edd8690c0c28d1f2ab.

Feng J, Shen ZJM, Zhan L (2007) Ranked items auctions and online advertisement. *Prod. Oper. Manag.* 16(4):510–522.

Ghosh A, McAfee P, Papineni K, Vassilvitskii S (2009) Bidding for representative allocations for display advertising. *Proc. 5th Internat. Workshop on Internet and Network Econom (WINE'09)*, 208–219 (Springer, Berlin).

Hojjat A, Turner J, Cetintas S, Yang J (2017) A unified framework for the scheduling of guaranteed targeted display advertising under reach and frequency requirements. *Oper. Res.* 65(2):289–313.

Kontogiorgis S (2000) Practical piecewise-linear approximation for monotropic optimization. *INFORMS J. Comput.* 12(4):324–340.

Lejeune M, Turner J (2019) Planning online advertising using Gini indices. *Oper. Res.* 67(5):1222–1245.

Li Z, Tang Q, Floudas CA (2012) A comparative theoretical and computational study on robust counterpart optimization: II. Probabilistic guarantees on constraint satisfaction. *Ind. Eng. Chem. Res.* 51(19):6769–6788.

Lu S, Zhu Y, Dukes A (2015) Position auctions with budget-constraints: Implications for advertisers and publishers. *Marketing Sci.* 34(6):897–905.

Market Research Future (2019) Data management platform (DMP) market research report – Forecast up to 2023. Accessed August, 2019, https://www.marketresearchfuture.com/reports/data-management-platform-market-4573.

McAfee RP, Papineni K, Vassilvitskii S (2013) Maximally representative allocations for guaranteed delivery advertising campaigns. *Rev. Econ. Des.* 17(2):83–94.

Miller BL, Wagner HM (1965) Chance constrained programming with joint constraints. *Oper. Res.* 13(6):930–945.

Najafi-Asadolahi S, Fridgeirsdottir K (2014) Cost-per-click pricing for display advertising. *Manufacturing Service Oper. Management* 16(4):482–497.

Nemirovski A, Shapiro A (2006) Convex approximations of chance constrained programs. *SIAM J. Optim.* 17(4):969–996.

Parkes DC, Sandholm T (2005) Optimize-and-dispatch architecture for expressive ad auctions. *Proc. 6th ACM Conf. Electronic Commerce (EC'05)* (ACM, New York).

Postek K, Ben-Tal A, den Hertog D, Melenberg B (2018) Robust optimization with ambiguous stochastic constraints under mean and dispersion information. *Oper. Res.* 66(3).

Shamsi D, Holtan M, R L, Ye Y (2014) Online allocation rules in display advertising. Working paper, arXiv preprint arXiv:1407.5710.

Turner J (2012) The planning of guaranteed targeted display advertising. *Oper. Res.* 60(1):18–33.

Vee E, Vassilvitskii S, Shanmugasundaram J (2010) Optimal online assignment with forecasts. *Proc. 11th ACM Conf. Electronic Commerce (EC'10)*, 109–118 (ACM, New York).

Walsh WE, Boutilier C, Sandholm T, Shields R, Nemhauser GL, Parkes DC (2010) Automated channel abstraction for advertising auctions. *Proc. 24th AAAI Conf. Artificial Intelligence*, 887–894 (Association for the Advancement of Artificial Intelligence, Atlanta, GA).

Yang J, Vee E, Vassilvitskii S, Tomlin J, Shanmugasundaram J, Anastasakos T, Kennedy O (2010) Inventory allocation for online graphical display advertising. Technical Report YL-2010-04, Yahoo! Labs, Sunnyvale, CA.

Zhang X, Feng J (2011) Cyclical bid adjustments in search-engine advertising. *Management Sci.* 57(9):1703–1719.

Zhu Y, Wilbur KC (2011) Hybrid advertising auctions. *Marketing Sci.* 30(2):249–273.

Zipkin PH (1980) Bounds for aggregating nodes in network problems. *Math. Programming* 19(1):155–177.

Zymler S, Kuhn D, Rustem B (2013) Distributionally robust joint chance constraints with second-order moment information. *Math. Programming* 137(1-2):167–198.