# An Asymptotically Tight Learning Algorithm for Mobile-Promotion Platforms

Zhichao Feng

Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University
fengzc10@gmail.com

Milind Dawande, Ganesh Janakiraman, Anyan Qi

Naveen Jindal School of Management, The University of Texas at Dallas
milind@utdalla.edu, ganesh@utdallas.edu, axq140430@utdallas.edu

**Abstract**

Operating under both supply-side and demand-side uncertainties, a mobile-promotion platform conducts advertising campaigns for individual advertisers. Campaigns arrive dynamically over time, which is divided into seasons; each campaign requires the platform to deliver a target number of mobile impressions from a desired set of locations over a desired time interval. The platform fulfills these campaigns by procuring impressions from publishers, who supply advertising space on apps, via real-time bidding on ad exchanges. Each location is characterized by its *win curve*, i.e., the relationship between the bid price and the probability of winning an impression at that bid. The win curves at the various locations of interest are initially unknown to the platform, and it learns them on the fly based on the bids it places to win impressions and the realized outcomes. Each acquired impression is allocated to one of the ongoing campaigns. The platform's objective is to minimize its total cost (the amount spent in procuring impressions and the penalty incurred due to unmet targets of the campaigns) over the time horizon of interest. Our main result is a *bidding and allocation policy* for this problem. We show that our policy is the best possible (asymptotically tight) for the problem using the notion of *regret* under a policy, namely the difference between the expected total cost under that policy and the optimal cost for the *clairvoyant problem* (i.e., one in which the platform has full information about the win curves at all the locations in advance): The regret under *any* policy is $\Omega(\sqrt{I})$, where $I$ is the number of seasons, and that under our policy is $\mathcal{O}(\sqrt{I})$. We demonstrate the performance of our policy through numerical experiments on a test bed of instances whose input parameters are based on our observations at a real-world mobile-promotion platform.

**Keywords:** *online advertising*, *learning*, *regret minimization*, *stochastic dynamic programming*

## 1 Introduction

Mobile advertising, i.e., advertising on mobile devices such as smart phones or tablets, has now emerged as the dominant form of online advertising, with consumers spending increasingly more time on these devices (eMarketer 2019). The mobile ad market in the U.S. is predicted to increase from $76 billion in 2018 to $113.21 billion in 2020, surpassing the combined advertising expenditure on all traditional media, including TV and radio (eMarketer 2018). Not surprisingly, the sizable business opportunities in the mobile ad industry have led to the emergence of a variety of providers who help advertisers display their ads on mobile devices. One such player is a mobile-promotion platform – prominent examples include Centro (http://www.centro.net/), Cidewalk (http://www.cidewalk.com/), ExactDrive

(http://www.exactdrive.com/) – that accepts advertising campaigns from individual advertisers and procures mobile impressions via ad exchanges from locations of their interest over their chosen time durations to fulfill these campaigns. The problem we study in this paper emerged from our interactions with Cidewalk, Inc. We begin by introducing the primary features of the problem.

**The Demand Side:** An *impression* refers to an advertising opportunity that arises on a mobile application (app) when an end-user interacts with the app. Each *season*, e.g., two weeks or a month, mobile-promotion platforms such as Cidewalk contract with individual advertisers to deliver a certain number of impressions from their desired set of locations (cities, zip codes, or even smaller customized regions) over their desired time intervals within that season. We refer to each such contract with an individual advertiser as a *campaign*.

**The Supply Side:** To deliver the required number of impressions for the accepted campaigns, the platform procures impressions from *publishers* (content owners), who supply advertising space on apps, via real-time bidding on *ad exchanges* such as DoubleClick and OpenX. At each location of interest, the arrival of impressions is uncertain and is characterized by a location-specific arrival probability. The outcomes of the platform's bids to acquire impressions are also uncertain; at each location, we refer to the relation between the bid price and the probability of winning an impression at that bid as the *win curve* at that location. If the platform fails to meet the total requirement of impressions for a campaign, then it incurs a penalty cost for each unmet impression. This penalty cost could represent a monetary payment from the platform to the advertiser, or correspond to a loss of goodwill. For each impression that becomes available from a desired location, the platform determines the bid price in real time and, if it wins that impression, allocates it to an ongoing campaign. The platform's objective is to minimize its total cost (i.e., the amount spent in procuring impressions and the penalty cost) over the time horizon of interest.

**The Learning Component:** The probability of winning an impression increases in the bid the platform places to acquire that impression. The platform does not know the win curves at the various locations of interest in advance and thus needs to learn them on the fly based on the bids it places to win impressions and the realized outcomes. At each location, we consider a general parametric win curve characterized by a vector of location-specific parameters that are initially unknown to the platform. Thus, our problem involves the platform's (1) dynamic bidding for impressions, (2) allocation of the acquired impressions to ongoing campaigns, and (3) learning, i.e., estimating the parameters of the win curves at the locations of interest.

**Overview of the Analysis:** For convenience of exposition, we first analyze a *"static"* setting of the platform's problem, where the information about all the campaigns, namely their respective time durations and the desired number of impressions, in each season is available to the platform at the start

of that season. Aside from notational simplicity, the static version shares many of the core features of the problem with the *"dynamic"* version, where campaigns arrive dynamically over time. Therefore, it is convenient to first present our analysis of the static setting and then use it to investigate the dynamic version. For both settings, we present a *bidding and allocation policy* and analyze its performance. The performance of a policy is measured using its *regret*, i.e., the difference between the expected total cost under that policy and the optimal cost of the *clairvoyant problem* (i.e., one in which the platform has full information in advance about the win curves at all the locations). In both scenarios, we derive a lower bound on the regret under any policy in terms of the number of seasons and also establish a matching upper bound on the regret under our policy. We also illustrate the performance of our policy numerically.

We now summarize our main results and first explain our contributions relative to three papers that are the closest to our work. Later, in Section 1.2, we review other related literature.

### 1.1 Our Contributions

To our knowledge, our work is the first to study a mobile-promotion platform's impression acquisition and allocation problem that involves dynamic bidding, allocation, and learning. The main outcomes of our analysis are bidding and allocation policies for both static and dynamic arrival of campaigns. We show that the regret under each of the two policies is $\mathcal{O}(\sqrt{I})$, where $I$ is the number of seasons. To establish a lower bound on the regret, we construct an instance for which the regret is $\Omega(\sqrt{I})$ under *any* policy. Thus, we obtain an asymptotically tight bound, namely $\Theta(\sqrt{I})$, on the regret. In addition, we analyze the special case where all impressions arrive from a single location and the win curve at that location satisfies the so-called "well-separated" condition defined in Broder and Rusmevichientong (2012), and propose a policy that achieves a $\Theta(\log I)$ regret. We also establish nuanced results with respect to other problem parameters such as the number of locations and the number of periods in each season. We demonstrate the performance of our policies through numerical experiments on a test bed of instances whose input parameters are inspired from our observations at Cidewalk.

As far as the advertising application is concerned, the problem studied in Aseri et al. (2017) is similar to ours, with the major difference being that they only analyze the clairvoyant problem, i.e., one in which full information about the win curves is available to the platform in advance. In contrast, our focus is on the platform's learning of the win curves at the various locations of interest. There are other relatively less-significant differences; e.g., in their model, the platform does not incur a penalty cost if it does not meet the requirement of a campaign. Instead, they impose a constraint that the requirement of each campaign be met with a high probability. The authors offer attractive policies with performance guarantees for both static and dynamic arrival of campaigns.

In terms of methodology, our work is closely related to two papers that incorporate learning of the demand distribution (i.e., the relationship between demand and price) – Broder and Rusmevichientong (2012) and den Boer and Zwart (2015). Since a detailed comparison of our setting and analysis with respect to these two papers would be appropriate only after our model and technical results have been presented, we relegate it to Appendix K.

We briefly mention some of the highlights of our technical analysis. While the platform needs to execute *both* the bidding for impressions and their allocation decisions on the fly, we leverage the characteristics of the advertising campaigns to quickly isolate the allocation decision and obtain an optimal allocation policy, and also simplify the platform's objective to a more-tractable one for the subsequent analysis of the dynamic bidding and learning decisions. The two types of uncertainties on the supply-side (namely, the uncertain arrival of impressions and the uncertain winning of impressions) and demand-side uncertainty (namely, the uncertain arrival of campaigns) result in a DP and a cost-to-go recursion that need to be analyzed to evaluate the performance of our bidding policy: the DP obtains the expected cost under the optimal bidding policy for the clairvoyant problem and the cost-to-go recursion computes the expected cost under an arbitrary bidding policy. We use the DP to derive an upper bound on the difference between the optimal bids under two arbitrary parameters of the win-curves. The cost-to-go recursion is used to obtain an upper bound on the regret under an arbitrary bidding policy; in turn, this upper bound helps us obtain an upper bound on the regret under our specific policy. Finally, under our bidding policy, the length of each exploration (exploitation) phase and the number of bids placed in each exploration phase are both random due to the uncertain arrival of impressions. We show that the regret under our policy essentially depends on the number of bids placed in each exploration (exploitation) phase instead of the length of each phase. In addition, we derive a uniform upper bound on the expected number of bids placed in each exploration phase.

In another highlight of our technical analysis, we establish two lower bounds on the regret under any policy in two settings. First, for the general problem, we show an $\Omega(\sqrt{I})$ lower bound on the regret under any policy. Similar to Broder and Rusmevichientong (2012), we apply the Kullback-Leibler (KL) divergence as a measure of the difference between two distributions to establish the lower bound. However, under two different values of the underlying parameters, Broder and Rusmevichientong (2012) compute the KL divergence of the distributions of the demands, while we use the KL divergence of the joint distributions of the outcomes of impression arrivals and the winning of impressions. Second, if we restrict our attention to the case where the total number of required impressions by the campaigns in each season is strictly less than the number of periods in a season, then we establish an $\Omega(I^{2/7})$ lower bound on the regret under any policy. In this case, we face an active capacity constraint, namely that the number of impressions assigned to each campaign cannot exceed its requirement, and need

to analyze the regret under this constraint.

Apart from the three papers discussed above, we now briefly review other related work.

### 1.2 Other Related Work

Being situated in the operations of a mobile-promotion platform, our work is naturally related to the literature on online display advertising. We refer the reader to Korula et al. (2015), Chen (2017), Agrawal et al. (2018), and Choi et al. (2019) for a comprehensive review of this literature. For brevity, here we focus on recent studies that address the learning of win curves or of the value of impressions by individual advertisers. Iyer et al. (2014) study bidding strategies of advertisers in repeated second-price auctions in which they learn, in a Bayesian fashion, their own distribution of the reward from winning an auction. The objective of each advertiser is to maximize the total expected payoff (rewards from winning auctions minus the bidding costs). They show that a mean field equilibrium exists in which it is optimal for each advertiser to bid truthfully. Zhang et al. (2014) propose a bidding strategy for an advertiser who first estimates the win curve (resp., the value of an impression) using least-square estimation (resp., Logistic regression) based on a training data set, and then bids to maximize the total expected value of winning impressions under a budget constraint. The effectiveness of the proposed bidding strategy is verified numerically. Balseiro and Gur (2019) consider the problem of advertisers bidding in repeated second-price auctions to maximize their respective total expected payoffs under budget constraints, without prior knowledge of the distributions of their own valuation of impressions or the distributions of the highest bids among their competitors. The authors propose an adaptive pacing bidding strategy, which dynamically adjusts the pace at which an advertiser depletes her budget using the realized expenditure in each period. They show that the strategy is asymptotically optimal and the regret under the strategy is $\mathcal{O}(\sqrt{N})$ if the advertiser's valuation and competitor's bids are independent and identically distributed, where $N$ is the number of auctions. When all the advertisers adopt such strategies, they characterize a regime under which these strategies constitute an approximate Nash equilibrium. Baardman et al. (2019) consider a multi-armed bandit problem of an advertiser deciding the portfolio of types (e.g., locations) of impressions to bid on in each of $T$ periods while learning the unknown revenue and cost of each type. In each period, the advertiser maximizes the expected total revenue subject to the budget of that period. The authors propose an optimistic-robust learning algorithm that achieves a regret of $\mathcal{O}(\log T)$.

There has been extensive work in recent years on demand learning in the Operations Management literature. For brevity, we avoid presenting an extensive review and limit ourselves to a few recent studies that develop results of the kind we obtain. Keskin and Zeevi (2014) consider a retailer selling multiple products over a time horizon of $T$ periods with unlimited inventory and assume a linear

demand function with unknown parameters. They develop a pricing policy, which is a variant of the well-known greedy iterated least squares policy, and show that the regret under their policy is $\Theta(\sqrt{T})$. Wang et al. (2014) consider a retailer selling a single product with finite inventory over a finite selling season, where the demand function is nonparametric and unknown to the retailer. For the problem where the demand function and initial inventory are scaled by $k > 0$, they propose a *learning-while-doing* pricing policy whose regret is $\mathcal{O}(\sqrt{k}\log^{4.5} k)$. Besbes and Zeevi (2012) consider a general network revenue management problem with multiple products and multiple limited resources. They consider an unknown and nonparametric demand function and show that the regret under their proposed policy is $\mathcal{O}(k^{(d+2)/(d+3)}\sqrt{\log k})$, where $d$ is the number of products. If the demand function is $s$-times differentiable, they propose another policy that reduces the regret to $\mathcal{O}(k^{(\frac{d}{s}+2)/(\frac{d}{s}+3)}\sqrt{\log k})$. Chen et al. (2019b) improve this upper bound by developing a *nonparametric self-adjusting control* that achieves a regret of $\mathcal{O}(k^{1/2+\epsilon}\log k)$ for any arbitrary small $\epsilon > 0$. Levi et al. (2015) consider a newsvendor problem with an unknown demand function. They analyze the performance of a sample-average approximation approach and show that the cumulative regret over $T$ periods is $\mathcal{O}(\log T)$. Chen et al. (2019a) study a joint pricing and inventory-replenishment problem with backorders over a planning horizon of $T$ periods where a retailer makes the replenishment and pricing decisions at the beginning of each period. They propose a nonparametric learning algorithm with regret $\mathcal{O}(\sqrt{T})$. Keskin et al. (2020) consider a utility company that dynamically sets electricity prices to serve $N$ customers over a time horizon of $T$ periods. The company initially knows neither the underlying cluster structure – induced by customer characteristics and exogenous factors – nor the consumption parameters in each cluster. The authors develop a data-driven policy, using spectral clustering and feature-based pricing, whose regret is $\mathcal{O}(\sqrt{NT})$ when all features are fully heterogeneous over time and customers. Keskin and Li (2021) study a dynamic pricing problem with unknown and time-varying heterogeneity in customers' preferences for quality. The expected number of market shifts is at most $n$ over $T$ periods in a Markovian market with unknown transition probabilities. The authors design a simple and practically implementable policy whose regret is $\mathcal{O}(\sqrt{n/T})$. For an excellent review of this literature, we refer the reader to den Boer (2015).

## 2 Static Campaign Arrivals

Recall that each advertising campaign is specified by its season (say, two weeks or a month), start time, end time, and the target number of impressions. The static setting assumes that, for all the campaigns within a season, this information is available at the beginning of that season. The platform bids on an ad exchange to win the impressions needed to fulfill the campaigns. We begin this section by precisely defining the static setting and stating our assumptions. Then, in Section 2.2, we derive an

optimal bidding and allocation policy for the full-information scenario (i.e., the clairvoyant problem). Finally, in Section 2.3, we derive an upper bound on the regret under *any* policy; this upper bound is used later to obtain an upper bound on the regret under our policy.

Each season consists of $T$ discrete time periods[1], where the length of a time period is sufficiently small so that at most one impression arrives in one time period over all locations of interest. Let $t \in \{1, \ldots, T\}$ denote the $t^{th}$ period of a season. In the $i^{th}$ season, information about all the campaigns that are to be executed in that season is available at the beginning of the season. Let $m_i$ denote the total number of campaigns in season $i$. For the $j^{th}$ campaign in the $i^{th}$ season, denoted by $(i, j)$, $j \in \{1, \ldots, m_i\}$, let $W_{i,j} > 0$ denote its target number of impressions and $\bar{t}_{i,j}, \underline{t}_{i,j} \in \{1, \ldots, T\}$ with $\bar{t}_{i,j} \leq \underline{t}_{i,j}$ denote its start and end time periods. Without loss of generality, we order these campaigns in increasing order of their end times, and let the end time of the last campaign in the season be $T$, i.e., $\underline{t}_{i,1} \leq \cdots \leq \underline{t}_{i,m_i} = T$. Let $\mathcal{C}_I$ denote the set of all the campaigns over the first $I$ seasons. That is,

$$\mathcal{C}_I = \bigcup_{i=1}^{I} \{(i,1), \ldots, (i, m_i)\}.$$

The concept of a season is defined as a practical "unit of time" to make it convenient for the platform to accept advertisement campaigns and for customers to specify their campaigns. Suppose the duration of a season is one month, with each season starting at the beginning of a month and finishing at the end of the month. In practice, if a customer (company) wants to engage with the platform over a long duration, then it requests the platform for a certain number of ad impressions per season (i.e., per month); for example, 30,000 ad impressions per season. Thus, the company signs a contract for several individual campaigns, each of duration one season. These individual campaigns are billed separately and ensure that the ad impressions are evenly distributed over the entire length of the engagement. In this manner, a long advertisement engagement is broken down into smaller campaigns that each has a duration of one season. Note that the length of a season is arbitrary (e.g., 2 weeks or 1 month or a quarter). In this sense, the assumption that all campaigns start and finish in the same season is a mild assumption. We will discuss the case where this assumption is not satisfied in Remark 4 in Section 3.2.

Let $\mathcal{L} = \{1, \ldots, L\}$ denote the set of locations; impressions acquired from any of these locations can be used to satisfy the requirement of any campaign. This is reasonable, for example, when the advertisers that the platform caters to belong to the same metropolitan area. In period $t$ of season $i$, denoted by $(i, t)$, an impression from location $l \in \mathcal{L}$ arrives with probability $q_l$. Let $\zeta_{i,t} = l$ if an impression arrives from location $l \in \mathcal{L}$ in period $(i, t)$ and $\zeta_{i,t} = 0$ if no impression arrives in that

[1]This is purely for expositional convenience. The analysis easily extends to the setting where the seasons are of different lengths.

period. Thus, $\mathbb{E}[\mathbb{1}\{\zeta_{i,t} = l\}] = q_l$. Let $b_{i,t}$ denote the platform's bid price in period $(i,t)$. If no impression arrives in period $(i,t)$, i.e., $\zeta_{i,t} = 0$, then no bid is placed and the bid price $b_{i,t} = 0$. If an impression arrives in period $(i,t)$, i.e., $\zeta_{i,t} \neq 0$, then the platform decides whether or not to place a bid. If no bid is placed, then the bid price $b_{i,t} = 0$. Otherwise, the platform chooses a bid price $b_{i,t} \in B = [b^{\min}, b^{\max}]$, where $b^{\max} > b^{\min} > 0$. Let $d_{i,t} = 1$ if the impression is won by bidding an amount $b_{i,t}$ at location $\zeta_{i,t}$, and $d_{i,t} = 0$ otherwise. Clearly, if the platform places no bid, i.e., $b_{i,t} = 0$, then no impression is won, i.e., $d_{i,t} = 0$. Let $p_l(\gamma_l, b_l)$ denote the *win curve* at location $l \in \mathcal{L}$, i.e., the probability of winning an impression that arrives from location $l$ by bidding an amount $b_l \in B$, where $\gamma_l = (\gamma_{l,1}, \ldots, \gamma_{l,n_l}) \in \Gamma_l \subset \mathbb{R}^{n_l}$ is a vector of $n_l$ parameters that characterize this distribution and $\Gamma_l$ is an open set. Then, $d_{i,t}$ is Bernoulli distributed with mean $p_{\zeta_{i,t}}(\gamma_{\zeta_{i,t}}, b_{i,t})$. The true value of $\gamma_l$ is unknown to the platform, and is denoted by $\gamma_l^{(0)}$; the platform learns this vector from the outcomes of the bids it places for winning impressions at location $l \in \mathcal{L}$. We assume that $\gamma_l^{(0)} \in \Gamma_l^{(0)}$, where $\Gamma_l^{(0)} \subset \Gamma_l$ is a compact and convex set. Let $\gamma = (\gamma_1, \ldots, \gamma_L)$, $\gamma^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_L^{(0)})$, $\Gamma = \Gamma_1 \times \cdots \times \Gamma_L$, and $\Gamma^{(0)} = \Gamma_1^{(0)} \times \cdots \times \Gamma_L^{(0)}$. The characterization of $p_l(\gamma_l, b_l)$ for $l \in \mathcal{L}$ is discussed in Section 2.1.

If an impression (from location $\zeta_{i,t}$) is acquired in period $(i,t)$, i.e., $d_{i,t} = 1$, then the platform needs to determine the campaign to which that impression is assigned. Let $a_{i,t}$ denote the allocation decision in period $(i,t)$, with $a_{i,t} = (i,j)$ indicating that if an impression arises in that period and is won, then it is allocated to campaign $(i,j)$. Let $c_{i,t,j}$ denote the number of *unmet* impressions for campaign $(i,j)$ at the beginning of period $(i,t)$. That is, at the beginning of period $(i,t)$, campaign $(i,j)$ needs a further $c_{i,t,j} \in \{0, \ldots, W_{i,j}\}$ impressions to be fulfilled. In time period $(i,t)$, we say that a campaign $(i,j)$ is *active* if an impression won in that period can be allocated to it; that is, $\bar{t}_{i,j} \leq t \leq \underline{t}_{i,j}$ and $c_{i,t,j} \geq 1$. Let $\mathcal{F}_{i,t} := \left\{(i,j) : \bar{t}_{i,j} \leq t \leq \underline{t}_{i,j}, c_{i,t,j} \geq 1, j \in \{1, \ldots, m_i\}\right\}$ denote the set of all active campaigns in time period $(i,t)$. Note that an impression won in period $(i,t)$ can only be allocated to an active campaign in that period, i.e., $a_{i,t} \in \mathcal{F}_{i,t}$.

We assume that the platform's bidding and allocation decisions in a time period only depend on the past history; specifically, the (i) arrival of impressions, (ii) the platform's bids, (iii) the realizations of the winning of impressions, (iv) the allocation of the impressions won to the various campaigns, and (v) the information about the campaigns. Let $h_{i,t}$ denote the history until the beginning of period $(i,t)$. Thus, we let

$$h_{i,t} := (\zeta_{\hat{i},\hat{t}}, b_{\hat{i},\hat{t}}, d_{\hat{i},\hat{t}}, a_{\hat{i},\hat{t}}, W_{\hat{i},j}, \bar{t}_{\hat{i},j}, \underline{t}_{\hat{i},j} : 1 \leq \hat{i} \leq i, (\hat{i},\hat{t}) < (i,t), 1 \leq j \leq m_{\hat{i}}), \tag{1}$$

where $(\hat{i},\hat{t}) < (i,t)$ if and only if $\hat{i} < i$ or $\hat{i} = i$ and $\hat{t} < t$. Let $\mathcal{H}_{i,t}$ denote the set of all possible histories until the beginning of period $(i,t)$.

The sequence of events in period $(i,t)$ is as follows: (i) the platform observes the history $h_{i,t}$

and makes the allocation decision[2] $a_{i,t}$; (ii) the impression arrival $\zeta_{i,t}$ is realized; (iii) the platform determines the bid price $b_{i,t}$; (iv) the binary outcome $d_{i,t}$, which indicates whether or not the platform wins the impression, is realized; (v) if the impression is won, i.e., $d_{i,t} = 1$, then the platform pays the bid $b_{i,t}$ and allocates the impression to campaign $a_{i,t}$. A *non-anticipating* (deterministic) policy $\pi$ is then defined as $\pi := (b^{\pi}_{i,t,l}(h_{i,t}), a^{\pi}_{i,t}(h_{i,t}) : i \in \{1,\dots,I\}, t \in \{1,\dots,T\}, l \in \mathcal{L}, h_{i,t} \in \mathcal{H}_{i,t})$, where $b^{\pi}_{i,t,l}(h_{i,t})$ is the bid price for an impression that arrives from location $l$ in period $(i,t)$ and $a^{\pi}_{i,t}(h_{i,t})$ is the campaign to which the impression won in that period is assigned under the history $h_{i,t} \in \mathcal{H}_{i,t}$. For notational convenience, we use $b^{\pi}_{i,t,l}$ (resp., $a^{\pi}_{i,t}$) to denote $b^{\pi}_{i,t,l}(h_{i,t})$ (resp., $a^{\pi}_{i,t}(h_{i,t})$) whenever no confusion arises in doing so. Let

$$x_{i,t} := (\zeta_{\hat{i},\hat{t}}, d_{\hat{i},\hat{t}}, W_{\hat{i},j}, \bar{t}_{\hat{i},j}, \underline{t}_{\hat{i},j} : 1 \le \hat{i} \le i, (\hat{i},\hat{t}) < (i,t), 1 \le j \le m_{\hat{i}}). \tag{2}$$

We show in Appendix A that for any $x_{i,t}$ and policy $\pi$, we can find the unique corresponding history $h^{\pi}_{i,t} = h^{\pi}_{i,t}(x_{i,t})$. Therefore, for notational brevity, we refer to $x_{i,t}$ as the history until the beginning of period $(i,t)$, and with slight abuse of notation, we use $b^{\pi}_{i,t,l}(x_{i,t})$ (resp., $a^{\pi}_{i,t}(x_{i,t})$) to denote $b^{\pi}_{i,t,l}(h^{\pi}_{i,t}(x_{i,t}))$ (resp., $a^{\pi}_{i,t}(h^{\pi}_{i,t}(x_{i,t}))$) in what follows. To further ease exposition, we drop the superscript $\pi$ in $h^{\pi}_{i,t}(x_{i,t})$ and denote it simply as $h_{i,t}(x_{i,t})$.

The platform's bidding cost is the cost it incurs in procuring the impressions. In period $(i,t)$, the bidding cost incurred under policy $\pi$ is $\sum_{l\in\mathcal{L}} b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t} = l\}d_{i,t}$. If the platform fails to fulfill a campaign (i.e., does not deliver the number of impressions needed to fulfill that campaign), then it incurs a penalty cost. For each campaign $(i,j)$, let $e_{i,j} \in [e^{\min}, e^{\max}]$ denote the unit penalty cost for each unmet impression[3]. Then, the penalty cost for campaign $(i,j)$ is

$$\left[W_{i,j} - \sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t} = l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t} = j\}\right]^{+} e_{i,j}.$$

Thus, the total expected cost (i.e., bidding cost plus penalty cost) under policy $\pi$ after $I$ seasons is

$$\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t} = l\}d_{i,t} + \sum_{j=1}^{m_i}\left[W_{i,j} - \sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t} = l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t} = j\}\right]^{+} e_{i,j}\right]. \tag{3}$$

The platform's goal is to obtain a bidding and allocation policy $\pi$ that minimizes its cumulative expected cost. For each location $l \in \mathcal{L}$, since the underlying vector of parameters $\gamma_l$ (that characterizes the distribution of $d_{i,t}$) is unknown, a policy should be careful in offering bids to adequately learn the unknown vector of parameters (i.e., compute a good estimate) at each location.

[2]Note that $a_{i,t}$ can be determined either before or after observing $\zeta_{i,t}$ and $d_{i,t}$.

[3]The unit penalty cost $e_{i,j}$ for each unmet impression includes both the opportunity cost of the revenue that could have been earned from winning an impression, as well as additional direct penalty (i.e., the direct monetary penalty that the platform pays advertisers for each unmet impression). In Remark 1 below, we specify the two parts of the unit penalty cost $e_{i,j}$ in formulating an equivalent profit-maximization problem for the platform.

It is immediate that, for any campaign $(i,j)$, it is optimal for the platform to not assign more than $W_{i,j}$ impressions to that campaign. We let $\Pi$ denote the set of all non-anticipating policies satisfying $\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t}=j\}\leq W_{i,j}$ a.s. for all $(i,j)\in\mathcal{C}_I$. Under any policy $\pi\in\Pi$, we can rewrite the total expected cost in (3) as

$$\begin{aligned}
&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}+\sum_{j=1}^{m_i}\left[W_{i,j}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t}=j\}\right]e_{i,j}\right] \qquad (4)\\
=&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\sum_{j=1}^{m_i}e_{i,j}\mathbb{1}\{a^{\pi}_{i,t}=j\}\right]+\sum_{i=1}^{I}\sum_{j=1}^{m_i}W_{i,j}e_{i,j}\\
=&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}e_{i,a^{\pi}_{i,t}}\right]+\sum_{i=1}^{I}\sum_{j=1}^{m_i}W_{i,j}e_{i,j}\\
=&\,\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b^{\pi}_{i,t,l}-e_{i,a^{\pi}_{i,t}})q_l d_{i,t}\right]+\sum_{i=1}^{I}\sum_{j=1}^{m_i}W_{i,j}e_{i,j}.
\end{aligned}$$

The third equality holds, since $\mathbb{E}[\mathbb{1}\{\zeta_{i,t}=l\}]=q_l$. Thus, the platform's problem can now be equivalently written as

$$\min_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b^{\pi}_{i,t,l}-e_{i,a^{\pi}_{i,t}})q_l d_{i,t}\right].$$

**Remark 1: (Profit Maximization Version)** Let $r_{i,j}$ denote the unit revenue the platform collects for each impression supplied to satisfy the demand of campaign $(i,j)$. Let $\hat{e}_{i,j}$ be the direct unit penalty cost for each unmet impression of campaign $(i,j)$ (i.e., the direct monetary penalty that the platform pays to the advertiser for each unmet impression of campaign $(i,j)$). Then, the platform's profit-maximization problem, over all policies $\pi\in\Pi$, is:

$$\begin{aligned}
&\max_{\pi\in\Pi}\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\sum_{j=1}^{m_i}\mathbb{1}\{a^{\pi}_{i,t}=j\}(r_{i,j}-b^{\pi}_{i,t,l})-\sum_{j=1}^{m_i}\left[W_{i,j}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t}=j\}\right]\hat{e}_{i,j}\right]\\
\Leftrightarrow&\max_{\pi\in\Pi}\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(r_{i,a^{\pi}_{i,t}}-b^{\pi}_{i,t,l})\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}-\sum_{j=1}^{m_i}\left[W_{i,j}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t}=j\}\right]\hat{e}_{i,j}\right]-\sum_{i=1}^{I}\sum_{j=1}^{m_i}r_{i,j}W_{i,j}\\
\Leftrightarrow&\min_{\pi\in\Pi}\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}+\sum_{j=1}^{m_i}\left[W_{i,j}-\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a^{\pi}_{i,t}=j\}\right](r_{i,j}+\hat{e}_{i,j})\right]. \qquad (5)
\end{aligned}$$

Comparing the objective (5) above to the cost-minimization objective (4) we defined earlier, note that the unit penalty cost $e_{i,j}$ for each unmet impression of campaign $(i,j)$ in (4) corresponds to $r_{i,j}+\hat{e}_{i,j}$ (i.e., the opportunity cost of the unit revenue that could have been earned from winning an impression plus the direct unit penalty cost for each unmet impression of campaign $(i,j)$) in (5). ■

Our analysis with respect to the penalty cost is organized as follows. In the remainder of this section and in Sections 3 and 4, we assume that the penalty cost for each unmet impression is the

same across different campaigns, denoted by $e$. This assumption is motivated by our observation of a mobile-promotion platform in practice: nearly all the customers of this platform are small- to medium-sized businesses, with similar valuations for an advertising opportunity. In Remark 5 in Section 3.2 and in Appendix I, we discuss the robustness of our results by analyzing a setting where the unit penalty cost for an unmet impression differs across campaigns. In Section 5, we numerically examine the behavior of the regret under our policy when the unit penalty cost for an unmet impression differs across campaigns.

Under the assumption of the same penalty cost across campaigns, the platform's problem can be written as in $(P)$ below.

$$\min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b_{i,t,l}^{\pi}-e)q_l d_{i,t}\right]. \tag{P}$$

**Remark 2:** It is clear from the above objective that, if $e \leq b^{\min}$, then not placing any bid (thus resulting in the objective function value of 0) is optimal for the platform, since it loses money by placing a bid. Therefore, we assume henceforth that $e > b^{\min}$. Also, for ease of exposition, we refer to the objective of problem $(P)$ as the expected cost of the platform. ■

**First-End-First-Serve (FEFS) allocation:** We first specify the campaign to which an impression won in period $(i,t)$ is assigned. Recall that $\mathcal{F}_{i,t}$ is the set of all active campaigns in time period $(i,t)$. It is easy to see that the following is an optimal allocation policy: In any period, allocate the impression won (if any) in that period to the active campaign that ends first, i.e., $a_{i,t}^{\pi} = (i, g_{i,t})$, where $g_{i,t} = \min_{(i,j)\in\mathcal{F}_{i,t}} j$. We refer to such an allocation policy as a FEFS policy[4] and formally note its optimality.

**Property:** Without loss of optimality, we can assume that the allocation policy is FEFS.

Note that if there are no active campaigns in period $(i,t)$, i.e., $\mathcal{F}_{i,t} = \emptyset$, then no bid should be placed and no impression is won in that period; thus, the allocation $a_{i,t}^{\pi}$ is of no consequence and can be chosen arbitrarily.

## 2.1 Win Curves

We now discuss our assumptions on the win curve $p_l(\gamma_l, b_l)$, i.e., the probability of winning an impression that arrives from location $l \in \mathcal{L}$ by bidding a price $b_l \in B$, with the vector of parameters $\gamma_l = (\gamma_{l,1}, \ldots, \gamma_{l,n_l}) \in \Gamma_l$ characterizing this distribution. Note that the function $p_l(\gamma_l, b_l)$ is defined on $\Gamma_l \times B$, and the true value of $\gamma_l$ (i.e., $\gamma_l^{(0)}$) is assumed to be in $\Gamma_l^{(0)} \subseteq \Gamma_l$. For $l \in \mathcal{L}$, we impose the following assumptions on the win curve $p_l(\gamma_l, b_l)$.

[4]For the setting where the unit penalty cost for an unmet impression differs across campaigns (see Remark 5 at the end of Section 3.2 and Appendix I), the FEFS allocation policy may not be optimal.

**Assumption 1** *For any $l \in \mathcal{L}$, $p_l(\gamma_l, b_l) \in C^2$ (twice continuously differentiable in $\gamma_l$ and $b_l$) for all $b_l \in B$ and $\gamma_l \in \Gamma_l$. Further, $p_l(\gamma_l, b_l)$ is log concave in $b_l$, $p_l(\gamma_l, b_l) \in (0,1)$ and $\frac{\partial p_l(\gamma_l, b_l)}{\partial b_l} > 0$ for all $b_l \in B$ and $\gamma_l \in \Gamma_l^{(0)}$.*

Under the above assumption, the probability of winning an impression from any location $l \in \mathcal{L}$ is bounded away from 0 and 1 on the bid interval $B$, and increases in the bid price $b_l \in B$. Many families of parametric win curves satisfy the above assumption (for appropriate choices of $\Gamma_l$, $\Gamma_l^{(0)}$ and $B$) including $p_l(\gamma_l, b_l) = \gamma_{l,1} + \gamma_{l,2} b_l$ (linear win curve), $p_l(\gamma_l, b_l) = \exp(\gamma_{l,1} + \gamma_{l,2} b_l)$ (exponential win curve), and $p_l(\gamma_l, b_l) = \frac{\exp(\gamma_{l,1} + \gamma_{l,2} b_l)}{1 + \exp(\gamma_{l,1} + \gamma_{l,2} b_l)}$ (logit win curve). Such linear, exponential, and logit forms have been discussed in Broder and Rusmevichientong (2012) and den Boer and Zwart (2015).

We also impose a statistical assumption. For each location $l \in \mathcal{L}$, let $Q_l^{\boldsymbol{b_l}, \gamma_l} : \{0,1\}^k \to [0,1]$ denote the probability distribution of the outcome $\boldsymbol{D} = (D_1, \cdots, D_k)$ of the winning of impressions for a given sequence of fixed bids $\boldsymbol{b_l} = (b_{l,1}, \cdots, b_{l,k}) \in B^k$. This distribution is represented by

$$Q_l^{\boldsymbol{b_l}, \gamma_l}(\boldsymbol{d}) = \prod_{\hat{k}=1}^{k} p_l(\gamma_l, b_{l,\hat{k}})^{d_{\hat{k}}} (1 - p_l(\gamma_l, b_{l,\hat{k}}))^{1 - d_{\hat{k}}},$$

where $\boldsymbol{d} \in \{0,1\}^k$ denotes an arbitrary realization of the random vector $\boldsymbol{D}$.

**Assumption 2** *(Statistical Assumption). For any $l \in \mathcal{L}$, there exist $k_l \in \mathbb{N}$ and a vector of exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}} = (\bar{b}_{l,1}, \cdots, \bar{b}_{l,k_l}) \in B^{k_l}$ such that the family of distributions $\{Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l} : \gamma_l \in \Gamma_l^{(0)}\}$ is identifiable, i.e., $\forall\ \gamma_l \neq \bar{\gamma}_l$, $\exists\ \boldsymbol{d} \in \{0,1\}^{k_l}$, s.t. $Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l}(\boldsymbol{d}) \neq Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \bar{\gamma}_l}(\boldsymbol{d})$. Moreover, the Fisher information matrix $\boldsymbol{I_l}(\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l)$, given by*

$$[\boldsymbol{I_l}(\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l)]_{u,v} = \mathbb{E}\left[ -\frac{\partial^2}{\partial \gamma_{l,u} \partial \gamma_{l,v}} \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l}(\boldsymbol{D}) \right],\ \textit{for } u, v = 1, \ldots, n_l, \tag{6}$$

*is positive definite.*

Assumption 2 is a common assumption (see, e.g., Besbes and Zeevi 2009 and Broder and Rusmevichientong 2012) and guarantees that we can estimate the vector of parameters $\gamma_l^{(0)}$ based on the observations of the impressions won at the exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}}$. As shown in the following examples, many parametric win curves satisfy Assumptions 1 and 2.

**Example 1 (linear win curve).** Let $B = [1/2, 1]$, $\Gamma_l^{(0)} = [1/8, 1/4] \times [1/3, 2/3]$, $\Gamma_l = (0, 7/24) \times (0, 17/24)$, and $p_l(\gamma_l, b_l) = \gamma_{l,1} + \gamma_{l,2} b_l$ for all $l \in \mathcal{L}$. It is straightforward to check that Assumption 1 is satisfied and $\{Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l} : \gamma_l \in \Gamma_l^{(0)}\}$ is identifiable for any $\bar{\boldsymbol{b}}_{\boldsymbol{l}} = (\bar{b}_{l,1}, \bar{b}_{l,2}) \in B^2$ with $\bar{b}_{l,1} \neq \bar{b}_{l,2}$. The Fisher information matrix is:

$$\boldsymbol{I_l}(\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l) = \frac{1}{p_l(\gamma_l, \bar{b}_{l,1})(1 - p_l(\gamma_l, \bar{b}_{l,1}))} \begin{pmatrix} 1 & \bar{b}_{l,1} \\ \bar{b}_{l,1} & \bar{b}_{l,1}^2 \end{pmatrix} + \frac{1}{p_l(\gamma_l, \bar{b}_{l,2})(1 - p_l(\gamma_l, \bar{b}_{l,2}))} \begin{pmatrix} 1 & \bar{b}_{l,2} \\ \bar{b}_{l,2} & \bar{b}_{l,2}^2 \end{pmatrix}.$$

It is easy to verify that the above matrix is positive definite. Thus, Assumption 2 is satisfied.

**Example 2 (exponential win curve).** Let $B = [1/2, 1]$, $\Gamma_l^{(0)} = [-3/2, -3/4] \times [1/3, 2/3]$, $\Gamma_l = (-2, -17/24) \times (0, 17/24)$, and $p_l(\gamma_l, b_l) = \exp(\gamma_{l,1} + \gamma_{l,2} b_l)$ for all $l \in \mathcal{L}$. Assumption 1 is satisfied and $\{Q_l^{\bar{\boldsymbol{b}}_l, \gamma_l} : \gamma_l \in \Gamma_l^{(0)}\}$ is identifiable for any $\bar{\boldsymbol{b}}_l = (\bar{b}_{l,1}, \bar{b}_{l,2}) \in B^2$ with $\bar{b}_{l,1} \neq \bar{b}_{l,2}$. The Fisher information matrix, which is positive definite (thus satisfying Assumption 2), is:

$$\boldsymbol{I}_l(\bar{\boldsymbol{b}}_l, \gamma_l) = \frac{p_l(\gamma_l, \bar{b}_{l,1})}{1 - p_l(\gamma_l, \bar{b}_{l,1})} \begin{pmatrix} 1 & \bar{b}_{l,1} \\ \bar{b}_{l,1} & \bar{b}_{l,1}^2 \end{pmatrix} + \frac{p_l(\gamma_l, \bar{b}_{l,2})}{1 - p_l(\gamma_l, \bar{b}_{l,2})} \begin{pmatrix} 1 & \bar{b}_{l,2} \\ \bar{b}_{l,2} & \bar{b}_{l,2}^2 \end{pmatrix}.$$

**Example 3 (logit win curve).** Let $B = [1/2, 1]$, $\Gamma_l^{(0)} = [-3/2, 3/2] \times [1/2, 3/2]$, $\Gamma_l = (-2, 2) \times (0, 2)$, and $p_l(\gamma_l, b_l) = \frac{\exp(\gamma_{l,1} + \gamma_{l,2} b_l)}{1 + \exp(\gamma_{l,1} + \gamma_{l,2} b_l)}$ for all $l \in \mathcal{L}$. Assumption 1 is satisfied and $\{Q_l^{\bar{\boldsymbol{b}}_l, \gamma_l} : \gamma_l \in \Gamma_l^{(0)}\}$ is identifiable for any $\bar{\boldsymbol{b}}_l = (\bar{b}_{l,1}, \bar{b}_{l,2}) \in B^2$ with $\bar{b}_{l,1} \neq \bar{b}_{l,2}$. Assumption 2 is also satisfied, since the Fisher information matrix

$$\boldsymbol{I}_l(\bar{\boldsymbol{b}}_l, \gamma_l) = p_l(\gamma_l, \bar{b}_{l,1})(1 - p_l(\gamma_l, \bar{b}_{l,1})) \begin{pmatrix} 1 & \bar{b}_{l,1} \\ \bar{b}_{l,1} & \bar{b}_{l,1}^2 \end{pmatrix} + p_l(\gamma_l, \bar{b}_{l,2})(1 - p_l(\gamma_l, \bar{b}_{l,2})) \begin{pmatrix} 1 & \bar{b}_{l,2} \\ \bar{b}_{l,2} & \bar{b}_{l,2}^2 \end{pmatrix}$$

is positive definite.

## 2.2 Optimal Bidding Policy for the Clairvoyant Problem (The Vector $\gamma$ is Known)

We now obtain an optimal bidding policy for the clairvoyant problem; i.e., the problem in which the vector of parameters of the win curves at all the locations $\gamma = (\gamma_1, \ldots, \gamma_L)$ is known, where $\gamma_l = (\gamma_{l,1}, \ldots, \gamma_{l,n_l})$ characterizes the win curve at location $l \in \mathcal{L} = \{1, \ldots, L\}$. It is important to note that, in our main optimization problem $(P)$, the learning of $\gamma$ occurs *across* seasons, in the sense that, in a given period, we can estimate $\gamma$ based on the bids placed in the past (including the bids placed before that period in the current season as well as those placed in all the previous seasons) and the corresponding outcomes. However, when $\gamma$ is known, no learning is needed. In this case, the optimization problem $(P)$ decouples into $I$ optimization problems, one for each season, since each campaign starts and ends within the same season. Therefore, it is sufficient to solve the problem corresponding to an individual season, say $i \in \{1, \cdots, I\}$.

Given that allocations are made in an FEFS manner, it is easy to see that the optimization problem for season $i$ can be written as the following DP, in which the state in any period $(i, t)$ is the number of unmet impressions at the beginning of that period for each campaign in the season; i.e., $(c_{i,t,1}, \cdots, c_{i,t,m_i})$. In the sequel, we drop the time index $(i, t)$ of $c_{i,t,j}$, $\mathcal{F}_{i,t}$, and $g_{i,t}$ when there is no ambiguity in doing so. Let $\boldsymbol{c_i} = (c_1, \cdots, c_{m_i})$. Let $V_{i,t}(\boldsymbol{c_i}; \gamma)$ denote the optimal cost-to-go function of the DP and let $b_{i,t,l}^*(\boldsymbol{c_i}; \gamma)$ denote the optimal bid price at location $l$ in period $(i, t)$ in state $\boldsymbol{c_i}$. Then, $V_{i,t}(\boldsymbol{c_i}; \gamma)$ satisfies the following recursion:

$$V_{i,t}(\boldsymbol{c_i}; \gamma)$$

$$= \min_{\substack{(b_1,\cdots,b_L):\\ b_l\in B_l, l\in\mathcal{L}}} \left\{ \begin{array}{l} \{\mathbb{1}\{\mathcal{F}\neq\emptyset\}\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)[b_l - e + V_{i,t+1}((c_1,\cdots,c_g-1,\cdots,c_{m_i});\gamma)] + \\ \left[1-\mathbb{1}\{\mathcal{F}\neq\emptyset\}\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)\right] V_{i,t+1}(\boldsymbol{c_i};\gamma) \end{array} \right\}$$

$$= \mathbb{1}\{\mathcal{F}\neq\emptyset\}\sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\gamma_l,b_l)\big[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i};\gamma)\big] + V_{i,t+1}(\boldsymbol{c_i};\gamma),$$

where $V_{i,T+1}(\boldsymbol{c_i};\gamma)=0$ and, for $\mathcal{F}\neq\emptyset$,

$$\Delta V_{i,t+1}(\boldsymbol{c_i};\gamma) = V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}((c_1,\cdots,c_g-1,\cdots,c_{m_i});\gamma).$$

For $\mathcal{F}\neq\emptyset$, the optimal bid price when an impression arrives from location $l\in\mathcal{L}$ is as follows:

$$b^*_{i,t,l}(\boldsymbol{c_i};\gamma) = \arg\min_{b_l\in B} p_l(\gamma_l,b_l)\left[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i};\gamma)\right]. \tag{7}$$

If $\mathcal{F}=\emptyset$, then no bid is placed, i.e., $b^*_{i,t,l}(\boldsymbol{c_i};\gamma)=0$ for all $l\in\mathcal{L}$. We show (in Lemma A.1, Appendix C) that $b^*_{i,t,l}(\boldsymbol{c_i};\gamma)$ is uniquely defined. The following technical assumption we make is similar in spirit to Assumption R4 in Chen et al. (2019b).

**Assumption 3** *For all $l\in\mathcal{L}$, $\gamma\in\Gamma^{(0)}$, $\mathcal{F}\neq\emptyset$, $1\le i\le I$, and $1\le t\le T$, the optimal bids $b^*_{i,t,l}(\boldsymbol{c_i};\gamma)\in(b^{\min},b^{\max})$, the interior of the bidding interval $B$.*

Under Assumption 3, we establish Lemma 1, which will be used later in Section 3.2 to obtain an upper bound on the regret under our policy. The proofs of the technical results are in the appendix.

Lemma 1 below shows that for any two vectors $\gamma$, $\hat{\gamma}$ of parameters of the win curves, the difference in the optimal bids under $\gamma$ and $\hat{\gamma}$ is $\mathcal{O}(\|\gamma-\hat{\gamma}\|)$, where $\|\cdot\|$ denotes the Euclidean norm.

**Lemma 1** *For all $l\in\mathcal{L}$, $\gamma,\hat{\gamma}\in\Gamma^{(0)}$, $\mathcal{F}\neq\emptyset$, $1\le i\le I$, and $1\le t\le T$, there exists a constant $K_1>1$ that is independent of $I$ and $T$, such that*

$$\left|b^*_{i,t,l}(\boldsymbol{c_i};\gamma) - b^*_{i,t,l}(\boldsymbol{c_i};\hat{\gamma})\right| \le (K_1)^T\|\gamma-\hat{\gamma}\|.$$

This concludes our analysis of the optimal policy when $\gamma$ is known. *In the remainder of Sections 2 and 3, we study the setting in which $\gamma$ is unknown.* Before we proceed with this, we define the performance metric we use (namely, regret) when $\gamma$ is unknown. This definition relies on the full-information setting studied above.

**Regret:** The performance of a policy is measured by its regret, which is defined as the expected increase in the platform's cost from not using the optimal bidding and allocation policy (under the true vector of parameters $\gamma^{(0)}$). The platform's optimal expected cost during season $i\in\{1,\cdots,I\}$, if

it knew $\gamma^{(0)}$, is $V_{i,1}((W_{i,1},\cdots,W_{i,m_i});\gamma^{(0)})$. Thus, the regret under a policy $\pi\in\Pi$ after $I$ seasons can be written as:

$$\text{Regret}(\pi,I;\gamma^{(0)})=\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b^{\pi}_{i,t,l}-e)q_l d_{i,t}\right]-\sum_{i=1}^{I}V_{i,1}((W_{i,1},\cdots,W_{i,m_i});\gamma^{(0)}).$$

### 2.3 An Upper Bound on the Regret Under Any Bidding and Allocation Policy

In this section, we derive an upper bound on the regret under an *arbitrary* bidding and FEFS allocation policy, say $\pi$. This will be useful in deriving an upper bound on the regret under our policy in Section 3.2. For any $\gamma\in\Gamma^{(0)}$, we first compute the expected cost in each season under policy $\pi$. Consider season $i$. Recall that $x_{i,1}$ is the history until the beginning of season $i$. Let $\hat{x}_{i,t}:=(\zeta_{i,\hat{t}},d_{i,\hat{t}}:1\le\hat{t}<t)$ denote the history in season $i$ until the beginning of period $(i,t)$. Then, the history until the beginning of period $(i,t)$ is $x_{i,t}=(x_{i,1},\hat{x}_{i,t})$; thus, the bid price for each location $l\in\mathcal{L}$ under policy $\pi$ is $b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t})$. Conditional on $x_{i,1}$, we compute the expected cost-to-go in season $i$ for state $\hat{x}_{i,t}$ under policy $\pi$, denoted by $V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)$, as follows.

For a given $\hat{x}_{i,t}$, we can find the associated number of unmet impressions of each campaign $(i,j)$ at the beginning of period $(i,t)$, denoted by $c_j(\hat{x}_{i,t})$; see Appendix B for details. Let $c(\hat{x}_{i,t}):=\sum_{j:\bar{t}_{i,j}\le t\le\underline{t}_{i,j}}c_j(\hat{x}_{i,t})$ denote the associated total number of unmet impressions at the beginning of period $(i,t)$ over all the ongoing campaigns in that period. Let $\boldsymbol{c_i}(\hat{x}_{i,t}):=(c_1(\hat{x}_{i,t}),\cdots,c_{m_i}(\hat{x}_{i,t}))$. $V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)$ satisfies the following recursion:

$$\begin{aligned}
&V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)=\\
&\mathbb{1}\{c(\hat{x}_{i,t})\ge 1\}\sum_{l\in\mathcal{L}}q_l p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t})-e+V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma)\right]+\\
&\sum_{l\in\mathcal{L}}q_l\left[1-\mathbb{1}\{c(\hat{x}_{i,t})\ge 1\}p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma)+\\
&\left(1-\sum_{l\in\mathcal{L}}q_l\right)V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma),
\end{aligned}$$

and $V^{\pi}_{i,T+1}(\cdot;x_{i,1},\gamma)=0$. Note that if $c(\hat{x}_{i,t})=0$, then there are no active campaigns, and thus no bid should be placed, i.e., $b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t})=0$ for all $l\in\mathcal{L}$.

Under policy $\pi$, let $X^{\pi}_{i,1}$ denote the random history until the beginning of season $i$ and $\hat{X}^{\pi}_{i,t}$ denote the random history in season $i$ until the beginning of period $(i,t)$. Then, the expected cost under policy $\pi$ after $I$ seasons is $\sum_{i=1}^{I}\mathbb{E}\left[V^{\pi}_{i,1}(\emptyset;X^{\pi}_{i,1},\gamma^{(0)})\right]$, and thus the regret under policy $\pi$ after $I$ seasons is:

$$\text{Regret}(\pi,I;\gamma^{(0)})=\sum_{i=1}^{I}\mathbb{E}\left[V^{\pi}_{i,1}(\emptyset;X^{\pi}_{i,1},\gamma^{(0)})\right]-\sum_{i=1}^{I}V_{i,1}((W_{i,1},\cdots,W_{i,m_i});\gamma^{(0)}).$$

Lemma 2 below provides an upper bound on the above regret.

**Lemma 2** *There exists a constant $K_0 > 0$ that is independent of $I$ and $T$, such that the regret under any bidding and FEFS allocation policy $\pi$ after $I$ seasons satisfies*

$$\textit{Regret}(\pi, I; \gamma^{(0)}) \ \leq K_0 \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} q_l \left(b^{\pi}_{i,t,l}(X^{\pi}_{i,1}, \hat{X}^{\pi}_{i,t}) - b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,t}); \gamma^{(0)})\right)^2\right].$$

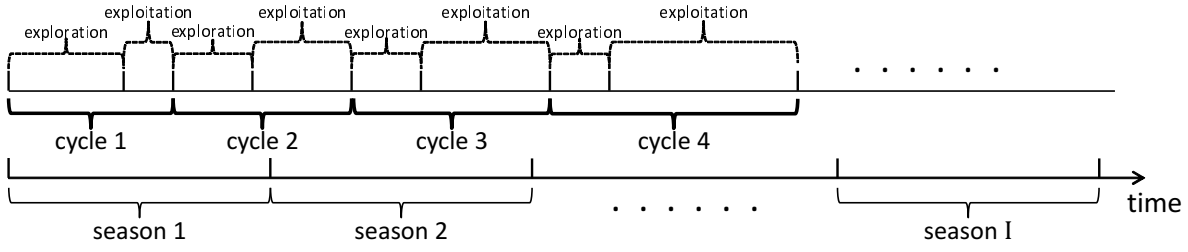
## 3 Analysis of Static Campaign Arrivals

In Section 3.1, we present our bidding and allocation policy for the setting where the vector of parameters $\gamma = (\gamma_1, \cdots, \gamma_L)$ of the win curves is unknown, where $\gamma_l = (\gamma_{l,1}, \cdots, \gamma_{l,n_l})$ characterizes the win curve at location $l \in \mathcal{L} = \{1, \cdots, L\}$. Section 3.2 establishes an $\mathcal{O}(\sqrt{I})$ upper bound on the regret under this policy. In Section 3.3, we establish a matching $\Omega(\sqrt{I})$ lower bound on the regret under *any* policy. For the special case where the total number of required impressions over all the campaigns in each season is strictly less than the number of periods in that season, we obtain an $\Omega(I^{2/7})$ lower bound on the regret under *any* policy.

### 3.1 Bidding and Allocation Policy

We refer to our bidding and allocation policy as BidAlloc. Recall that allocations are made in an FEFS manner. The formal description of BidAlloc follows.

Figure 1: The basic structure of BidAlloc, our bidding and allocation policy.




**Policy** BidAlloc

**Input:** For each location $l \in \mathcal{L}$, the exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}} = (\bar{b}_l^1, \cdots, \bar{b}_l^{k_l})$; see Section 2.1.

**Bidding:** The bidding policy operates in cycles, with each cycle consisting of an exploration phase and an exploitation phase. Figure 1 shows the basic structure of the policy. Let $s$ denote the index of a cycle, starting with $s = 1$. We refer to the infinite sequence $\bar{\boldsymbol{b}}_{\boldsymbol{l}}^{\boldsymbol{\infty}} = (\bar{b}_l^1, \cdots, \bar{b}_l^{k_l}, \bar{b}_l^1, \cdots, \bar{b}_l^{k_l}, \cdots)$,

which iteratively repeats the sequence $\bar{\boldsymbol{b}}_{\boldsymbol{l}}$, as the *exploration sequence* for location $l \in \mathcal{L}$. Let $\Upsilon_l$ denote a counter for the exploration sequence for location $l$.

Initialize $\Upsilon_l = 0$ for all $l \in \mathcal{L}$. We now describe the two phases of an arbitrary cycle $s$, $s \geq 1$.

- *Exploration Phase of Cycle $s$*: Consider any period $(i, t)$ in this phase.
  - If there are no active campaigns or if no impression arrives, then no bid is placed.
  - Otherwise, for an impression that arrives from location $l \in \mathcal{L}$, the exploration counter for $l$ increases by 1, that is, $\Upsilon_l = \Upsilon_l + 1$. The bid placed is the element in position $\Upsilon_l$ of the exploration sequence $\bar{\boldsymbol{b}}_{\boldsymbol{l}}^{\boldsymbol{\infty}}$.

  This phase concludes whenever we have $\Upsilon_l \geq sk_l$ for all $l \in \mathcal{L}$; that is, every bid in $\bar{\boldsymbol{b}}_{\boldsymbol{l}}$ has been "explored" at least $s$ times cumulatively from cycle 1.

  At the end of this exploration phase, for each location $l \in \mathcal{L}$, consider the first $sk_l$ realized outcomes of the bids placed at that location. Note that these observations correspond to the placing of the exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}} = (\bar{b}_l^1, \cdots, \bar{b}_l^{k_l})$ repeatedly $s$ times. For $1 \leq \hat{s} \leq s$, let $\boldsymbol{D_l}(\hat{s}) = (D_l^1(\hat{s}), \cdots, D_l^{k_l}(\hat{s}))$ denote the corresponding outcomes when the exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}}$ are placed for the $\hat{s}^{th}$ time. Let $\hat{\gamma}_l(s)$ denote the maximum-likelihood estimate (MLE)[5] based on these $sk_l$ observations; that is,

$$\hat{\gamma}_l(s) = \arg\max_{\gamma_l \in \Gamma_l^{(0)}} \prod_{\hat{s}=1}^{s} Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l}(\boldsymbol{D_l}(\hat{s})).$$

- *Exploitation Phase of Cycle $s$*: Consider any period $(i, t)$ in this phase.
  - If there are no active campaigns or if no impression arrives, then no bid is placed.
  - Otherwise, for an impression that arrives from location $l \in \mathcal{L}$, place the bid $b^*_{i,t,l}(\boldsymbol{c_i}; \hat{\gamma}(s))$, computed by Equation (7) using the vector of estimates $\hat{\gamma}(s) = (\hat{\gamma}_1(s), \cdots, \hat{\gamma}_L(s))$.

  This phase concludes when a total of $Ls$ bids are placed in this phase over all locations. Then, the exploration phase for cycle $s + 1$ begins.

**Allocation:** If an impression is acquired in a period, then allocate it to the active campaign that ends first.

Lemma 3, below guarantees that, after a sufficient number of exploration cycles, our estimate $\hat{\gamma}(s)$ is guaranteed to be close to $\gamma^{(0)}$, in the following precise sense:

[5]Note that both the MLE formulation (e.g., den Boer and Zwart (2015) and Broder and Rusmevichientong (2012)) and the Bayesian approach (e.g., Sunar et al. (2021), Qi et al. (2017), Harrison and Sunar (2015), and Harrison et al. (2012)) are common modeling approaches for parametric learning problems. While our problem can also be modeled with the Bayesian approach, our use of the MLE formulation is just a matter of modeling choice.

### Mean-Squared Errors for MLE Based on IID Samples

**Lemma 3** *There exists a constant $K_{mle} > 0$ that is independent of $I$ and $T$, such that, for any $s \geq 1$, the vector of the maximum-likelihood estimates $\hat{\gamma}(s) = (\hat{\gamma}_1(s), \cdots, \hat{\gamma}_L(s))$ after $s$ exploration phases satisfies:*

$$\mathbb{E}\left[\left\|\hat{\gamma}(s) - \gamma^{(0)}\right\|^2\right] \leq \frac{K_{mle}}{s}.$$

We note that the basis for Lemma 3 is a fundamental result – Theorem 36.3 in Borovkov (1998) – on the convergence of maximum-likelihood estimators.

## 3.2 Upper Bound on the Regret Under BidAlloc

The main result of this section is Theorem 1, which establishes an upper bound on the regret under the policy BidAlloc.

**Theorem 1** *Under Assumptions 1, 2, and 3, the policy* BidAlloc *satisfies*[6]

$$\textit{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq K_3\sqrt{I},$$

*where $K_3 = K_2(K_1)^{2T}\sqrt{T}$ for constants $K_1$ and $K_2$ that are independent of $I$ and $T$.*

**Proof of Theorem 1:** Let $\hat{\pi}$ denote the policy BidAlloc. Let $S_0 = \lceil\sqrt{2IT/L}\rceil$. Then, the total duration of the first $S_0$ cycles is at least $IT$ periods because the total number of bids placed during the exploitation phases of $S_0$ cycles is $\sum_{s=1}^{S_0} Ls = LS_0(S_0+1)/2 \geq IT$. Let $X^{\hat{\pi}}$ denote the random history over $I$ seasons under policy $\hat{\pi}$. For any realized history $x$, let $\mathcal{A}_{s_1}(x)$ (resp., $\mathcal{A}_{s_2}(x)$) denote the collection of periods belonging to the exploration (resp., exploitation) phase of cycle $s$ in which there are active campaigns. Then, the regret of policy $\hat{\pi}$ after $I$ seasons satisfies:

$$\begin{aligned}
&\sum_{i=1}^{I} \mathbb{E}\left[V_{i,1}^{\hat{\pi}}(\emptyset; X_{i,1}^{\hat{\pi}}, \gamma^{(0)}) - V_{i,1}((W_{i,1}, \cdots, W_{i,m_i}); \gamma^{(0)})\right] \\
&\leq K_0\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b_{i,t,l}^{\hat{\pi}}(X_{i,1}^{\hat{\pi}}, \hat{X}_{i,t}^{\hat{\pi}}) - b_{i,t,l}^{*}(\boldsymbol{c_i}(\hat{X}_{i,t}^{\hat{\pi}}); \gamma^{(0)})\right)^2\right] \text{ (using Lemma 2)} \\
&= K_0\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t} = l\}\left(b_{i,t,l}^{\hat{\pi}}(X_{i,1}^{\hat{\pi}}, \hat{X}_{i,t}^{\hat{\pi}}) - b_{i,t,l}^{*}(\boldsymbol{c_i}(\hat{X}_{i,t}^{\hat{\pi}}); \gamma^{(0)})\right)^2\right] \\
&= K_0\sum_{s=1}^{S_0}\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t} = l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_1}(X^{\hat{\pi}})\}\left(b_{i,t,l}^{\hat{\pi}}(X_{i,1}^{\hat{\pi}}, \hat{X}_{i,t}^{\hat{\pi}}) - b_{i,t,l}^{*}(\boldsymbol{c_i}(\hat{X}_{i,t}^{\hat{\pi}}); \gamma^{(0)})\right)^2\right] +
\end{aligned}$$

[6] We also show that the regret is $\mathcal{O}(T)$ in Lemma A.9 and $\mathcal{O}(\sqrt{T}\log^2(T))$ under the setting defined in Theorem A.1; see Remark 3 for more details.

$$K_0 \sum_{s=1}^{S_0} \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_2}(X^{\hat{\pi}})\}\left(b^{\hat{\pi}}_{i,t,l}(X^{\hat{\pi}}_{i,1},\hat{X}^{\hat{\pi}}_{i,t}) - b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\hat{\pi}}_{i,t});\gamma^{(0)})\right)^2\right]. \quad (8)$$

The first equality holds since $\mathbb{E}[\mathbb{1}\{\zeta_{i,t}=l\}] = q_l$ and $\zeta_{i,t}$ is independent of $\hat{X}^{\hat{\pi}}_{i,t}$ and $X^{\hat{\pi}}_{i,1}$. The second equality holds because for those periods with no active campaigns, we have $b^{\hat{\pi}}_{i,t,l}(X^{\hat{\pi}}_{i,1},\hat{X}^{\hat{\pi}}_{i,t}) = b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\hat{\pi}}_{i,t});\gamma^{(0)}) = 0$. We will show that the first (resp., second) expectation in each cycle in (8) is bounded from above by a constant $\hat{K} > 0$ (resp., $\check{K} = LK_{mle}(K_1)^{2T} > 0$) that is independent of $I$ and $s$. Thus, we have

$$\begin{aligned}\sum_{i=1}^{I}\mathbb{E}\left[V^{\hat{\pi}}_{i,1}(\emptyset;X^{\hat{\pi}}_{i,1},\gamma^{(0)}) - V_{i,1}((W_{i,1},\cdots,W_{i,m_i});\gamma^{(0)})\right] &\le K_0 S_0(\hat{K}+\check{K}) \quad (9)\\ &= \left\lceil\sqrt{\frac{2IT}{L}}\right\rceil K_0(\hat{K}+\check{K})\\ &\le \left(\sqrt{\frac{2IT}{L}}+1\right)K_0(\hat{K}+\check{K})\\ &\le K_3\sqrt{I},\end{aligned}$$

where $K_3 = K_2(K_1)^{2T}\sqrt{T}$ for $K_2 = (\sqrt{\frac{2}{L}}+1)K_0(\hat{K}+LK_{mle})$.

We now derive an upper bound on the first expectation in (8):

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_1}(X^{\hat{\pi}})\}\left(b^{\hat{\pi}}_{i,t,l}(X^{\hat{\pi}}_{i,1},\hat{X}^{\hat{\pi}}_{i,t}) - b^{*}_{i,t,l}(c(\hat{X}^{\hat{\pi}}_{i,t});\gamma^{(0)})\right)^2\right]\\ &\le (b^{\max}-b^{\min})^2\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_1}(X^{\hat{\pi}})\}\right]\\ &\le (b^{\max}-b^{\min})^2\sum_{l\in\mathcal{L}}\frac{k_l}{q_l}.\end{aligned}$$

Thus, we let $\hat{K} = (b^{\max}-b^{\min})^2\sum_{l\in\mathcal{L}}\frac{k_l}{q_l}$.

Next, we derive an upper bound on the second expectation in (8). Let $\mathcal{X}^{\hat{\pi}}$ denote the set of all possible histories under policy $\hat{\pi}$. For any $x\in\mathcal{X}^{\hat{\pi}}$, let $\Pr(x)$ denote the probability of $x$ to be the realized history under policy $\hat{\pi}$, $\zeta_{i,t}(x)$ denote the realized impression arrival in period $(i,t)$, and $\hat{\gamma}(s;x)$ denote the realized estimates based on $x$. Then, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_2}(X^{\hat{\pi}})\}\left(b^{\hat{\pi}}_{i,t,l}(X^{\hat{\pi}}_{i,1},\hat{X}^{\hat{\pi}}_{i,t}) - b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\hat{\pi}}_{i,t});\gamma^{(0)})\right)^2\right]\\ &= \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_2}(X^{\hat{\pi}})\}\left(b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\hat{\pi}}_{i,t});\hat{\gamma}(s)) - b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\hat{\pi}}_{i,t});\gamma^{(0)})\right)^2\right]\\ &\le (K_1)^{2T}\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} \mathbb{1}\{\zeta_{i,t}=l\}\mathbb{1}\{(i,t)\in\mathcal{A}_{s_2}(X^{\hat{\pi}})\}\left\|\hat{\gamma}(s)-\gamma^{(0)}\right\|^2\right]\end{aligned}$$

$$= (K_1)^{2T} \sum_{x \in \mathcal{X}^{\hat{\pi}}} \Pr(x) \sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \mathbb{1}\{\zeta_{i,t}(x) = l\} \mathbb{1}\{(i,t) \in \mathcal{A}_{s_2}(x)\} \left\| \hat{\gamma}(s;x) - \gamma^{(0)} \right\|^2.$$

The inequality holds by Lemma 1.

Notice that for any given history $x$, the total number of bids placed in the exploitation phase of cycle $s$ is $\sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \mathbb{1}\{\zeta_{i,t}(x) = l\} \mathbb{1}\{(i,t) \in \mathcal{A}_{s_2}(x)\}$. Recall that in the exploitation phase of cycle $s$, the number of bids placed is $Ls$. Thus, for any $x$, we have

$$\sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \mathbb{1}\{\zeta_{i,t}(x) = l\} \mathbb{1}\{(i,t) \in \mathcal{A}_{s_2}(x)\} = Ls, \text{ and}$$

$$\begin{aligned}
&(K_1)^{2T} \sum_{x \in \mathcal{X}^{\hat{\pi}}} \Pr(x) \sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \mathbb{1}\{\zeta_{i,t}(x) = l\} \mathbb{1}\{(i,t) \in \mathcal{A}_{s_2}(x)\} \left\| \hat{\gamma}(s;x) - \gamma^{(0)} \right\|^2 \\
&= (K_1)^{2T} \sum_{x \in \mathcal{X}^{\hat{\pi}}} \Pr(x) Ls \left\| \hat{\gamma}(s;x) - \gamma^{(0)} \right\|^2 \\
&= (K_1)^{2T} Ls \mathbb{E}\left[ \left\| \hat{\gamma}(s) - \gamma^{(0)} \right\|^2 \right] \\
&\leq (K_1)^{2T} Ls \frac{K_{mle}}{s} = (K_1)^{2T} L K_{mle}.
\end{aligned}$$

The inequality holds by Lemma 3. Thus, let $\check{K} = (K_1)^{2T} L K_{mle}$. This completes the proof of (9). ■

**Remark 3:** We also obtain an upper bound on the regret with respect to $T$. Specifically, we show that the regret under any policy is $\mathcal{O}(T)$, and that the constant in the definition of $\mathcal{O}(T)$ is independent of the number of locations $L$. Further, under the special case where all the impressions arrive from a single location, whose win curve is $p(\gamma, b) = \exp(\gamma(b - e))$, and the start and end times of the campaigns in each season are ordered in the same way (i.e., the campaigns end in the order of their arrival), we show in Theorem A.1 that the regret under our policy is $\mathcal{O}(\sqrt{T} \log^2(T))$. We refer the reader to Appendix G for more details. ■

**Remark 4:** Our analysis assumes that a campaign starts and finishes in the same season. Without this assumption (i.e., when campaigns can start and end in different seasons), there is no "decomposition" of campaigns across time, and hence a general analysis becomes intractable. However, under the special case where all impressions arrive from a single location, whose win curve is $p(\gamma, b) = \exp(\gamma(b-e))$, and the start and end times of campaigns are ordered in the same way, Theorem A.1 (which we discussed in Remark 3 above) helps us establish that the regret under our policy is $\mathcal{O}(\sqrt{I} \log^2(I))$ (Theorem A.2). We refer the reader to Appendix G for more details.

**Remark 5 (Analysis of the Regret Under Other Allocation Policies):** Consider the setting where the unit penalty cost for an unmet impression differs across campaigns, and hence the FEFS

allocation policy may not be optimal. Consider the following general class of non-anticipating (deterministic) allocation policies: In each season, the allocation decision in each period of that season is deterministically defined based on the history in that season until the beginning of that period. More precisely, the active campaign to which an impression acquired in period $(i,t)$ from location $l$ is assigned (i.e., the allocation decision $a_{i,t,l}$) is deterministically defined based on the history in season $i$ until the beginning of that period, i.e., $\hat{x}_{i,t}$. Let $\Phi$ denote the set of all such allocation policies. Given an allocation policy[7] in $\Phi$, the platform only needs to decide its bidding policy. In Theorem A.5 of Appendix I, we show that, under the given allocation policy, the regret under our bidding policy remains $\mathcal{O}(\sqrt{I})$, where $I$ is the total number of seasons. Thus, our learning algorithm is effective under *any* allocation policy in $\Phi$.

To establish Theorem A.5, we formulate a DP and a cost-to-go recursion that both need to be analyzed to evaluate the performance of our bidding policy: the DP defines the expected cost under the optimal bidding policy for the clairvoyant problem and the cost-to-go recursion defines the expected cost under an arbitrary bidding policy. Note that when the unit penalty costs differ across campaigns and we are given an arbitrary allocation policy in $\Phi$, the DP and the cost-to-go recursion are significantly different from the ones in our analysis of Theorem 1, which only applies to the FEFS allocation policy. We use the DP to derive an upper bound on the difference between the optimal bids under two arbitrary parameters of the win-curves (Lemma A.14). The DP and the cost-to-go recursion are both used to obtain an upper bound on the regret under an arbitrary bidding policy (Lemma A.15); in turn, this upper bound helps us obtain an upper bound on the regret under our specific policy. Finally, using Lemmas A.14 and A.15, we show that the regret under our policy is $\mathcal{O}(\sqrt{I})$. We refer the reader to Appendix I and Appendix J for more details. ■

### 3.3 Lower Bounds on the Regret Under Any Policy

In Theorem 2 below, we obtain an $\Omega(\sqrt{I})$ lower bound on the regret by constructing an instance of problem $(P)$ that satisfies Assumptions 1, 2, and 3, and whose worst-case regret is $\Omega(\sqrt{I})$ under *any* policy. For the setting where the total number of required impressions by the campaigns in each season is strictly less than the number of periods in a season, we establish an $\Omega(I^{2/7})$ lower bound on the regret under *any* policy in Theorem 3.

**Theorem 2** *Consider the following instance of problem $(P)$: $b^{\min} = 5/8$, $b^{\max} = 11/8$, $e = 2$, $T = 1$ and $I \geq 2$. There is only one location, i.e., $L = 1$. In each period, an impression arrives with probability $q > 0$. The probability of winning an arriving impression under a bid price $b \in B =$*

[7]Examples of allocation policies in the set $\Phi$ include the FEFS policy and the policy that allocates an acquired impression to the active campaign with the highest ratio of penalty cost to the remaining duration of the campaign.

$[b^{\min}, b^{\max}]$ *is* $p(\gamma, b) = 1/2 - \gamma + \gamma b$*, where* $\gamma \in \Gamma^{(0)} = [1/3, 1]$ *and* $\Gamma^{(0)} \subset \Gamma = (0, 4/3)$*. There is one campaign in each season and the required number of impressions is no less than the number of periods in the season, so that the target of the campaign can never be exceeded. Then, the following statements hold:*

*(i) Assumptions 1, 2, and 3 are satisfied for this instance.*

*(ii) For any policy* $\pi$*, there exists a true parameter* $\gamma^{(0)} \in \Gamma^{(0)}$ *such that the regret after* $I$ *seasons satisfies:*

$$\textit{Regret}(\pi, I; \gamma^{(0)}) \geq \frac{q}{2(24^3)}\sqrt{I}.$$

The intuition is as follows. For the instance defined in Theorem 2, note that when $b = 1$, we have $p(\gamma, b) = 1/2$, regardless of the value of the underlying parameter $\gamma$. Thus, bidding at this amount does not help us gain information about the value of $\gamma$; thus, $b = 1$ is an *uninformative* bid. To learn $\gamma$, we need to place some bids away from the uninformative bid. However, the uninformative bid is the optimal bid when $\gamma = 1/2$; thus, placing bids away from the optimal bid increases the total expected cost. This leads to the $\Omega(\sqrt{I})$ lower bound on the regret in Theorem 2. The proof of Theorem 2 is provided in Appendix D.1.

In the instance we used in Theorem 2, the required number of impressions by a campaign in a season is no less than the number of periods in the season. den Boer and Zwart (2015) study the dynamic pricing of multiple products and obtain an $\Omega(\log I)$ lower bound on the regret under any policy when the initial inventory of each season is strictly less than the number of periods in a season. They show that their problem satisfies an endogenous learning property and propose a pricing policy which achieves an $\mathcal{O}(\log^2(I))$ upper bound on the regret. However, in our context, if the total number of required impressions by the campaigns in each season is strictly less than the number of periods in a season, then due to the random arrival of impressions, the endogenous learning property does not hold and the $\mathcal{O}(\log^2(I))$ upper bound on the regret cannot be achieved. Specifically, in this case, Theorem 3 below establishes an $\Omega(I^{2/7})$ lower bound on the regret under any policy.

**Theorem 3** *Consider the following instance of problem* $(P)$*:* $b^{\min} = 5/8$*,* $b^{\max} = 11/8$*,* $e = 2$*,* $T = 2$ *and* $I \geq 2$*. There is only one location, i.e.,* $L = 1$*. In each period, an impression arrives with probability* $q = K_4 I^{-1/7}$ *for a constant* $0 < K_4 < \sqrt{\frac{31}{24\sqrt{6}}}$ *that is independent of* $I$*. The probability of winning an arriving impression under a bid price* $b \in B = [b^{\min}, b^{\max}]$ *is* $p(\gamma, b) = 1/2 - \gamma + \gamma b$*, where* $\gamma \in \Gamma^{(0)} = [1/3, 1]$ *and* $\Gamma^{(0)} \subset \Gamma = (0, 4/3)$*. There is one campaign in each season whose required number of impressions is one. Then, the following statements hold:*

*(i) Assumptions 1, 2, and 3 are satisfied for this instance.*

*(ii) For any policy $\pi$, there exists a true parameter $\gamma^{(0)} \in \Gamma^{(0)}$ such that the regret after $I$ seasons satisfies:*

$$Regret(\pi, I; \gamma^{(0)}) \geq \frac{1}{8}\left[\frac{\sqrt{K_4}}{96\sqrt{6}} - \frac{(K_4)^{5/2}}{124}\right]^2 I^{2/7}.$$

We discuss the intuition. Consider an arbitrary season. Notice that, in the instance defined in Theorem 3, the required number of impressions by the campaign in the season (namely, 1) is strictly less than the number of periods in the season (namely, 2). Following the same argument as in the intuition of Theorem 2, if no impression arrives in the first period of the season, then the optimal bid in the second period is the uninformative bid $b = 1$ when $\gamma = 1/2$. We need to place some bids away from the uninformative bid, which increases the total expected cost. If the arrival probability $q$ of impressions is reasonably small, then with high probability, no impression arrives in the first period of the season. However, $q$ should not be too small. This is because if no impression arrives in both periods of the season, then no bids will be placed in the season and the regret in that season will be zero. By choosing an appropriate value of $q$ (namely, $q = K_4 I^{-1/7}$), we derive an $\Omega(I^{2/7})$ lower bound on the regret under any policy in Theorem 3. The proof of Theorem 3 is provided in Appendix D.2.

## 4 Dynamic Campaign Arrivals

We now consider the setting in which campaigns arrive dynamically. To distinguish the notation from that of the static model, we use superscript $D$ to denote "Dynamic", where necessary. At most one campaign can arrive in a period, and a campaign arriving in period $(i,t)$ (if any) is assumed to arrive at the beginning of that period. For any $w \geq 0$ and $\tau \geq 1$, let $\lambda_{i,t}^{w,\tau}$ denote the probability that a campaign which requires $w$ impressions and ends in period $(i,\tau)$ arrives in period $(i,t)$; let $\mathcal{W}$ denote the set of all possible values of the tuple $(w,\tau)$. All the other details of the dynamic setting are the same as those in the static setting defined in Section 2. Note that, similar to the static setting, we can also restrict our attention to FEFS allocation policies without loss of optimality.

The flow of our analysis in this section is similar to that for the static setting in Sections 2 and 3. As with the static model, we first obtain an optimal bidding policy for the clairvoyant problem; i.e., the problem in which the vector of parameters of the win curves at all the locations $\gamma = (\gamma_1, \cdots, \gamma_L)$ is known, where $\gamma_l = (\gamma_{l,1}, \cdots, \gamma_{l,n_l})$ characterizes the win curve at location $l \in \mathcal{L} = \{1, \cdots, L\}$.

### 4.1 Optimal Bidding Policy for the Clairvoyant Problem

For the clairvoyant problem, we specify the optimal bidding policy in each season by formulating the following DP. Consider season $i$. Let $c_{i,t,\tau}$ denote the number of unmet impressions at the beginning of period $(i,t)$ for all season-$i$ campaigns that start in or before period $(i,t)$ and end in period $(i,\tau)$; the

time index $(i,t)$ of $c_{i,t,\tau}$ will be dropped when there is no ambiguity in doing so. Let $\boldsymbol{c} = (c_1, \cdots, c_T)$, $g_t(\boldsymbol{c}) = \min\{\tau : \tau \geq t, c_\tau \geq 1\}$ and $\eta_t(\boldsymbol{c}) = \sum_{\tau=t}^{T} c_\tau$. Note that if $\eta_t(\boldsymbol{c}) \geq 1$, i.e., there are active campaigns in period $(i,t)$, then the impression won in that period should be allocated to one of the active campaigns that ends in period $(i, g_t(\boldsymbol{c}))$. Let $\boldsymbol{e_\tau}$ denote a unit vector of length $T$ whose $\tau^{th}$ component equals to 1. The optimal cost-to-go $V_{i,t}^D(\boldsymbol{c};\gamma)$ at the beginning of period $(i,t)$ satisfies the following recursion:

$$V_{i,t}^D(\boldsymbol{c};\gamma) =$$

$$\sum_{(w,\tau)\in\mathcal{W}} \lambda_{i,t}^{w,\tau} \min_{\substack{(b_1,\ldots,b_L):\\ b_l\in B, l\in\mathcal{L}}} \left\{ \begin{array}{l} \mathbb{1}\left\{\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) \ \geq 1\right\} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l) \left[b_l - e + V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau} - \boldsymbol{e_{g_t(\boldsymbol{c}+w\boldsymbol{e_\tau})}};\gamma)\right] \ + \\ \left(1 - \mathbb{1}\left\{\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) \ \geq 1\right\} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l)\right) V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) \end{array} \right\}$$

$$= \sum_{(w,\tau)\in\mathcal{W}} \lambda_{i,t}^{w,\tau} \left( \mathbb{1}\left\{\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) \ \geq 1\right\} \sum_{l\in\mathcal{L}} q_l \min_{b_l\in B} p_l(\gamma_l, b_l) \left[b_l - e - \Delta V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma)\right] + V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) \right),$$

where $\Delta V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) = V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) - V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau} - \boldsymbol{e_{g_t(\boldsymbol{c}+w\boldsymbol{e_\tau})}};\gamma)$ for all $\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) \ \geq 1$ and $V_{i,T+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) = 0$ for all $\eta_T(\boldsymbol{c}+w\boldsymbol{e_\tau}) \ \geq 0$.

Let $b_{i,t,l}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma)$ denote the optimal bid price at location $l \in \mathcal{L}$ in period $(i,t)$. If $\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) = 0$, then no bid should be placed, i.e., $b_{i,t,l}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) = 0$ for all $l \in \mathcal{L}$. Otherwise, the optimal bid at location $l$ is:

$$b_{i,t,l}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) = \arg\min_{b_l\in B} p_l(\gamma_l, b_l) \left[b_l - e - \Delta V_{i,t+1}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma)\right].$$

As in the static setting, $b_{i,t,l}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma)$ is uniquely defined (Lemma A.1 in Appendix C). Analogous to Assumption 3 and Lemma 1 for the static model, we assume that the optimal bids lie strictly in the interior of $B$ and establish Lemma 4, which will be used later in Section 4.3 to derive an upper bound on the regret under our policy.

**Assumption 4** *For all $\gamma \in \Gamma^{(0)}$, $l \in \mathcal{L}$, $1 \leq i \leq I$, $1 \leq t \leq T$, and $\eta_t(\boldsymbol{c}+w\boldsymbol{e_\tau}) \geq 1$, $b_{i,t,l}^D(\boldsymbol{c}+w\boldsymbol{e_\tau};\gamma) \in (b^{\min}, b^{\max})$.*

Under Assumption 4, we show that for any two vectors $\gamma$, $\hat{\gamma}$ of parameters of the win curves, the difference in the optimal bids under $\gamma$ and $\hat{\gamma}$ is $\mathcal{O}(\|\gamma - \hat{\gamma}\|)$.

**Lemma 4** *For all $l \in \mathcal{L}$, $\gamma, \hat{\gamma} \in \Gamma^{(0)}$, $\eta_t(\boldsymbol{c}) \geq 1$, $1 \leq i \leq I$, and $1 \leq t \leq T$, there exists a constant $K_5 > 1$ that is independent of $I$ and $T$, such that*

$$\left|b_{i,t,l}^D(\boldsymbol{c};\gamma) - b_{i,t,l}^D(\boldsymbol{c};\hat{\gamma})\right| \leq (K_5)^T \|\gamma - \hat{\gamma}\|.$$

The proof of Lemma 4 is similar to that of Lemma 1, and is therefore omitted for brevity.

Next, we compute the expected cost under an arbitrary policy $\pi$ when $\gamma$ is unknown and derive an upper bound on the regret under that policy; along with Lemma 4, this bound will also be used to bound the regret under our policy in Section 4.3.

### 4.2 An Upper Bound on the Regret Under Any Bidding and Allocation Policy

Let $w_{i,t}$ (resp., $\tau_{i,t}$) denote the target number of impressions (resp., end time) of the campaign that arrives at the beginning of period $(i,t)$. Let $x_{i,t}^D$ denote the history before period $(i,t)$. That is,

$$x_{i,t}^D := \big(\zeta_{\hat{i},\hat{t}}, d_{\hat{i},\hat{t}}, w_{\hat{i},\hat{t}}, \tau_{\hat{i},\hat{t}} : (\hat{i},\hat{t}) < (i,t)\big).$$

Note that the history $x_{i,t}^D$ is comprised of the history before season $i$ (i.e., $x_{i,1}^D$) and the history within season $i$, denoted by $\hat{x}_{i,t}^D := (\zeta_{i,\hat{t}}, d_{i,\hat{t}}, w_{i,\hat{t}}, \tau_{i,\hat{t}} : 1 \le \hat{t} < t)$, i.e., $x_{i,t}^D = (x_{i,1}^D, \hat{x}_{i,t}^D)$. Thus, for an arbitrary policy $\pi$, the bid price $b_{i,t,l}^{\pi,D}(x_{i,1}^D, \hat{x}_{i,t}^D, w_{i,t}, \tau_{i,t})$ in period $(i,t)$ at location $l \in \mathcal{L}$ depends on $w_{i,t}$, $\tau_{i,t}$, and the history $x_{i,t}^D = (x_{i,1}^D, \hat{x}_{i,t}^D)$.

For a given $\hat{x}_{i,t}^D$, let $c_\tau(\hat{x}_{i,t}^D)$ denote the total number of unmet impressions at the beginning of period $(i,t)$ for all the campaigns that arrived before period $(i,t)$ and end in period $(i,\tau)$. Let $\boldsymbol{c}(\hat{x}_{i,t}^D) = (c_1(\hat{x}_{i,t}^D), \cdots, c_T(\hat{x}_{i,t}^D))$. Recall that $\eta_t(\boldsymbol{c}) = \sum_{\tau=t}^T c_\tau$ and $\eta_t(\boldsymbol{c}) \ge 1$ if and only if there are active campaigns in period $(i,t)$. Conditional on $x_{i,1}^D$, the expected cost-to-go $V_{i,t}^{\pi,D}(\hat{x}_{i,t}^D; x_{i,1}^D, \gamma)$ in season $i$ under policy $\pi$ satisfies:

$$\begin{aligned}
&V_{i,t}^{\pi,D}(\hat{x}_{i,t}^D; x_{i,1}^D, \gamma) \\
&= \sum_{(w,\tau)\in\mathcal{W}} \lambda_{i,t}^{w,\tau} \left\{ \begin{array}{l}
\mathbb{1}\left\{\eta_t(\boldsymbol{c}(\hat{x}_{i,t}^D) + w\boldsymbol{e_\tau}) \ge 1\right\} \sum_{l\in\mathcal{L}} q_l p_l \left(\gamma_l, b_{i,t,l}^{\pi,D}(x_{i,1}^D, \hat{x}_{i,t}^D, w, \tau)\right) \left[ \begin{array}{l} b_{i,t,l}^{\pi,D}(x_{i,1}^D, \hat{x}_{i,t}^D, w, \tau) - e + \\ V_{i,t+1}^{\pi,D}((\hat{x}_{i,t}^D, (l,1,w,\tau)); x_{i,1}^D, \gamma) \end{array} \right] + \\
\sum_{l\in\mathcal{L}} q_l \left[1 - \mathbb{1}\left\{\eta_t(\boldsymbol{c}(\hat{x}_{i,t}^D) + w\boldsymbol{e_\tau}) \ge 1\right\} p_l \left(\gamma_l, b_{i,t,l}^{\pi,D}(x_{i,1}^D, \hat{x}_{i,t}^D, w, \tau)\right)\right] V_{i,t+1}^{\pi,D}((\hat{x}_{i,t}^D, (l,0,w,\tau)); x_{i,1}^D, \gamma) + \\
\left(1 - \sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}^{\pi,D}((\hat{x}_{i,t}^D, (0,0,w,\tau)); x_{i,1}^D, \gamma)
\end{array} \right\},
\end{aligned}$$

and $V_{i,T+1}^{\pi,D}(\cdot; x_{i,1}^D, \gamma) = 0$. Note that if $\eta_t(\boldsymbol{c}(\hat{x}_{i,t}^D) + w\boldsymbol{e_\tau}) = 0$, then no bid is placed, i.e.,

$$b_{i,t,l}^{\pi,D}(x_{i,1}^D, \hat{x}_{i,t}^D, w, \tau) = 0 \text{ for all } l \in \mathcal{L}.$$

Let $X_{i,1}^{\pi,D}$ (resp., $\hat{X}_{i,t}^{\pi,D}$) denote the random history before season $i$ (resp., within season $i$, before period $(i,t)$) under policy $\pi$. Then, the expected cost under policy $\pi$ in season $i$ is $\mathbb{E}\left[V_{i,1}^{\pi,D}(\emptyset; X_{i,1}^{\pi,D}, \gamma^{(0)})\right]$, and the regret of policy $\pi$ after $I$ seasons is

$$\text{Regret}^D(\pi, I; \gamma^{(0)}) = \sum_{i=1}^I \mathbb{E}\left[V_{i,1}^{\pi,D}(\emptyset; X_{i,1}^{\pi,D}, \gamma^{(0)})\right] - \sum_{i=1}^I V_{i,1}^D(\mathbf{0}; \gamma^{(0)}).$$

Analogous to Lemma 2 in the static setting, the following result derives an upper bound on the above regret.

**Lemma 5** *Let $W_{i,t}$ ($\mathcal{T}_{i,t}$) denote the random number of target impressions (resp., end time) of the campaign that arrives in period $(i,t)$. Then, there exists a constant $K_0 > 0$ that is independent of $I$ and $T$, such that the regret under any policy $\pi$ after $I$ seasons satisfies*

$$\begin{aligned}&\textit{Regret}^D(\pi, I; \gamma^{(0)}) \leq \\ &K_0 \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} q_l \left(b_{i,t,l}^{\pi,D}\left(X_{i,1}^{\pi,D}, \hat{X}_{i,t}^{\pi,D}, W_{i,t}, \mathcal{T}_{i,t}\right) - b_{i,t,l}^{D}\left(\boldsymbol{c}(\hat{X}_{i,t}^{\pi,D}) + W_{i,t}\boldsymbol{e}_{\boldsymbol{\mathcal{T}_{i,t}}}; \gamma^{(0)}\right)\right)^2\right].\end{aligned}$$

The proof of Lemma 5 is similar to that of Lemma 2, and is therefore omitted for brevity.

### 4.3 Policy DynBid

We refer to our bidding policy under dynamic campaign arrivals as DynBid. This policy simply modifies the BidAlloc policy under static campaign arrivals as follows: In the exploitation phase of each cycle $s$, for an impression that arrives from location $l \in \mathcal{L}$, place the bid $b_{i,t,l}^{D}(\boldsymbol{c} + w\boldsymbol{e}_{\boldsymbol{\tau}}; \hat{\gamma}(s))$ instead of the bid $b_{i,t,l}^{*}(\boldsymbol{c_i}; \hat{\gamma}(s))$ under the BidAlloc policy.

**Upper Bound on the Regret of DynBid**: Theorem 4 below establishes an upper bound on the regret under the policy DynBid.

**Theorem 4** *Under Assumptions 1, 2, and 4, the policy* DynBid *satisfies*[8]

$$\textit{Regret}^D(\textsc{DynBid}, I; \gamma^{(0)}) \leq K_7\sqrt{I},$$

*where $K_7 = K_6(K_5)^{2T}\sqrt{T}$ for constants $K_5$ and $K_6$ that are independent of $I$ and $T$.*

The proof of Theorem 4 is similar to that of Theorem 1, and is therefore omitted for brevity.

**Lower Bounds on the Regret Under Any Policy**: The regret under any policy is $\Omega(\sqrt{I})$ (resp., $\Omega(I^{2/7})$) when the total number of required impressions by campaigns in each season is no less than (resp., strictly less than) the number of periods in the season, since the problem instance described in Theorem 2 (resp., Theorem 3) also serves as a special case of the setting of dynamic campaign arrivals, where exactly one campaign arrives at the beginning of each season.

**Remark 6:** Consider the special case where all impressions arrive from a single location and the win curve at that location satisfies the so-called "well-separated" condition defined in Broder and

[8] As in Remark 3, under dynamic campaign arrivals too, we can show that the regret is $\mathcal{O}(T)$ and $\mathcal{O}(\sqrt{T}\log^2(T))$ under conditions similar to that in Theorem A.1 in Appendix G. The details are omitted for brevity.

Rusmevichientong (2012). Under this condition, the uninformative bid is precluded; thus, the platform can exploit at the optimal bid based on the estimated win curve, and at the same time passively learn the parameters of the win curve. For a policy that is similar to the "greedy" policy presented in Section 4.2 of Broder and Rusmevichientong (2012), we show in Appendix H that the regret over $I$ seasons is $\Theta(\log I)$. ■

## 5 Numerical Analysis

In this section, we numerically illustrate the performance of our policy on a realistic setup that is based on our observations at Cidewalk. Section 5.1 describes our test bed. Section 5.2 discusses the approximations we use in our numerical computations for the dynamic programs in our learning algorithm. In Section 5.3, we illustrate the rate of the regret under our policy with respect to the number of seasons ($I$), the number of periods in each season ($T$), and the number of locations ($L$). We also consider the more-general setting where the unit penalty costs and the desired sets of geographical locations (from which impressions are sought) differ across campaigns, and illustrate the rate of the regret under our policy with respect to the number of seasons. In Section 5.4, we numerically decompose the total regret under our policy to assess how much of it is due to the approximations used in our computations.

### 5.1 Test Bed

In our test bed, the information regarding the geographical locations and the win curves is obtained from data made available by Cidewalk while the choices of the other parameters are based on our observations at that company.

In our base setting, each season consists of two weeks and there are $10^6$ time periods in each week. Thus, $T = 2 \times 10^6$, which implies that the duration of each time period is about 0.6 seconds. We consider three locations from the Boston area; these are indexed by $l = 1, 2, 3$ and corresponding to zip codes 02110, 02114, and 02116, respectively. The monetary unit in the data below is 0.1 cents. The win curve at location $l$ is $p_l(\gamma_l, b_l) = \frac{\exp(\gamma_{l,1}+\gamma_{l,2}b_l)}{1+\exp(\gamma_{l,1}+\gamma_{l,2}b_l)}; l = 1, 2, 3$, where $b_l \in B = [0.3, 8]$; the true values of the parameters in the win curves are $(\gamma_{1,1}, \gamma_{1,2}) = (-2.281, 0.705), (\gamma_{2,1}, \gamma_{2,2}) = (-2.192, 1.042)$, $(\gamma_{3,1}, \gamma_{3,2}) = (-1.905, 0.876)$, and $\Gamma_l^{(0)} = [-8, -0.1] \times [0.1, 8]$. Note that the true values of the parameters in the win curves are unknown to the platform in advance. The duration of a campaign is either one week or two weeks, and each campaign requires $80,000$ impressions. In each time period, an impression arrives from location $l \in \{1, 2, 3\}$ with probability 0.01. In the base case, the penalty cost of each unmet impression is 12.5; this cost is the same across different campaigns, as is observed in the status quo at Cidewalk. Campaigns can arrive at the beginning of each week.

Consider an arbitrary season: At most one campaign can arrive at the beginning of each week. At the beginning of the first week, either a one-week or a two-week campaign can arrive; each of these two events occurs with probability 0.45. At the beginning of the second week, a one-week campaign arrives with probability 0.9.

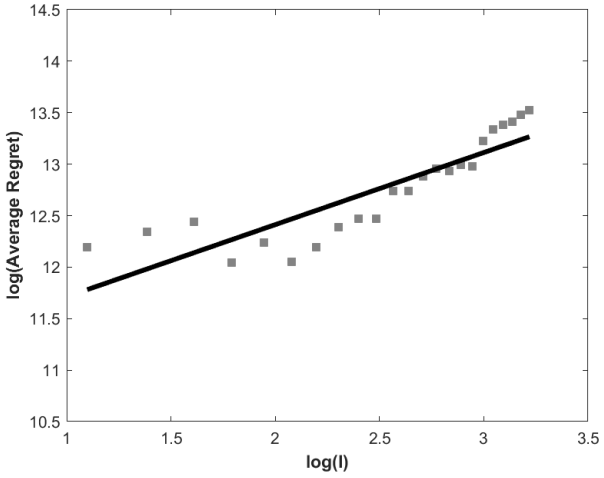
We also consider a generalized setting where the unit penalty cost and the desired set of locations may differ across campaigns: (i) the penalty cost of each unmet impression can take two values: 10 and 15. For a campaign arriving at the beginning of the first week, if the duration of a campaign is one week (resp., two weeks), then the unit penalty cost is 10 (resp., 15). For a campaign arriving at the beginning of the second week, the duration is one week and the unit penalty cost can either be 10 or 15. (ii) Each campaign requires impressions from two out of the three locations; accordingly, there are three desired sets of locations: $\{1,2\}$, $\{1,3\}$, and $\{2,3\}$. Thus, there are six possible campaign-types at the beginning of each week. For an arbitrary season, at the beginning of the first week, at most one campaign belonging to one of the six types can arrive; each of these six events occurs with probability 0.15. Similarly, at the beginning of the second week, at most one single-week campaign belonging to one of the six campaign types can arrive; each of these six events occurs with probability 0.15. All other details are the same as in the base case.

### 5.2 Approximating the Dynamic Programs in the Learning Algorithm

Recall that the clairvoyant problem in Section 4.1 is a DP with a multi-dimensional state space and its optimal solution is used as a benchmark to compute the regret under our policy. In our numerical study, instead of solving the DP optimally, which would require sophisticated and industry-strength software, we solve a convex optimization problem whose optimal objective value is a *lower bound* on the optimal cost of the clairvoyant problem. Also, recall from Section 4.3 that at the beginning of each exploitation phase, to compute the optimal bid based on the latest estimates, we need to solve a DP with a multi-dimensional state space. Similar to the convex optimization problem to approximate the DP of the clairvoyant problem, we define and solve a convex optimization problem instead of the DP in the exploitation phase. The optimal solution of this problem is used to obtain the bid price and allocation decision during the exploitation phase under our policy. Clearly, the expected cost under these bid prices and allocation decisions is greater than the optimal cost-to-go of the DP. Thus, through these two convex optimization problems, the regret we compute in our numerical study is an *upper bound* on the true regret under our policy.

The technical developments of the two convex optimization problems are in Appendix L.

Figure 2: Behavior of the regret with respect to the number of seasons, $I$.




### 5.3 Behavior of the Regret with Respect to $I$, $T$, and $L$

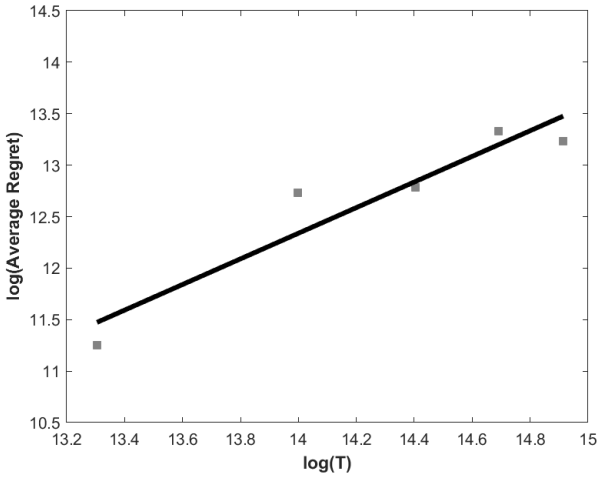
We first examine the behavior of the regret with respect to the number of seasons, $I$. For each value of $I$, we compute the average regret under our policy over 80 simulations under the base setting defined in Section 5.1. Figure 2 plots the logarithm[9] of the average regret as a function of the logarithm of the number of seasons, i.e., $\log(I)$, and also displays the corresponding best-fit line. Recall from Section 5.2 that we, in fact, compute an upper bound on the true regret under our policy. Also, recall from Theorem 4 that the true regret under our policy is $\mathcal{O}(\sqrt{I})$. Thus, the slope of the best-fit line would be close to 0.5 had we computed the true regret by optimally solving the DPs in our learning algorithm. Instead, since the average regret we plot here is an upper bound on the true regret, the slope of the best-fit line in Figure 2 is higher (approximately 0.70).

Next, we examine the regret under our policy with respect to the number of periods in each season ($T$) and the number of geographical locations ($L$), by fixing $I = 10$ and varying $T$ or $L$. The values of all the other parameter remain the same as in the base setting defined in Section 5.1. We vary the number of periods in each week from $3 \times 10^5$ to $1.5 \times 10^6$, in increments of $3 \times 10^5$. Thus, since each season consists of two weeks, the number of periods in a season, $T \in \{6 \times 10^5, 1.2 \times 10^6, \cdots, 3 \times 10^6\}$. Accordingly, we vary the required number of impressions by a campaign proportionately from $2.4 \times 10^4$ to $1.2 \times 10^5$, in increments of $2.4 \times 10^4$. For each value of $T$, we compute the average regret over 50 simulations. Figure 3 plots the logarithm of the average regret under our policy versus the logarithm of the number of periods, i.e., $\log(T)$, and also the corresponding best-fit line. Recall from Remark 3 (and Lemma A.9 in Appendix G) that the regret under any policy is $\mathcal{O}(T)$ and note that the average

[9]Throughout our computations, logarithm refers to the natural logarithm, i.e., to the base $e$.

regret we plot in Figure 3 is an upper bound on the true regret. Therefore, the slope (approximately 1.24) of the best-fit line is higher than the value of 1 suggested by our theoretical analysis. To examine

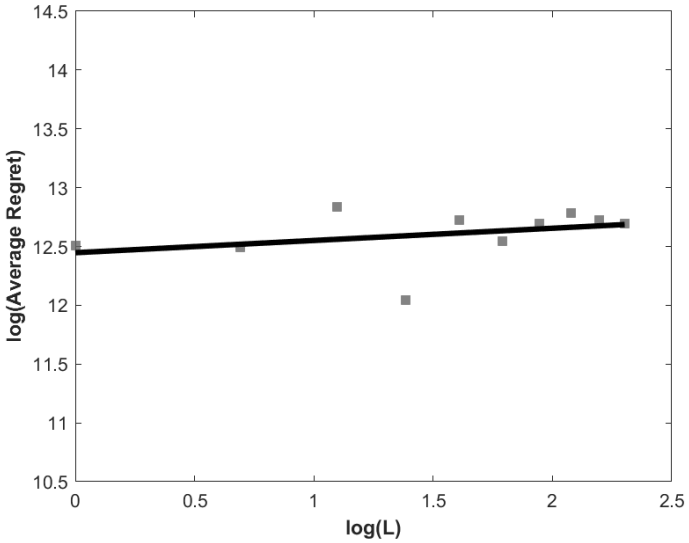
Figure 3: Behavior of the regret with respect to the number of time periods, $T$, in each season.




the behavior of the regret under our policy with respect to $L$ (the number of geographical locations from where impressions are sought), we vary $L$ from 1 to 10 with $(\gamma_{l,1}, \gamma_{l,2}) = (-2.281, 0.705)$ for $l \in \{1, 2, \cdots, 10\}$. For each value of $L$, we compute the average regret over 50 simulations. Figure 4 shows the logarithm of the average regret under our policy as a function of $\log(L)$. Recall from Remark 3 (and Lemma A.9 in Appendix G) that there exists an upper bound on the regret under any policy that is independent of $L$. In line with that analysis, the slope of the best-fit line in Figure 4 is close to 0 (approximately 0.10).

We also examine the behavior of the regret with respect to the number of seasons, $I$, under the generalized setting (defined in Section 5.1), where the unit penalty costs and the desired sets of locations possibly differ across campaigns. For each value of $I$, we compute the average regret under our policy over 80 simulations. Figure 5 plots the logarithm of the average regret under our policy with respect to the logarithm of the number of seasons, i.e., $\log(I)$. Note that the FEFS property (Section 2) no longer holds and the allocation decisions become significantly more complicated under the generalized setting for two reasons. First, since the unit penalty costs differ across campaigns, the allocation decision – i.e., the campaign to which an acquired impression is assigned – depends not only on the end times of the ongoing campaigns, but also on the unit penalty costs of those campaigns. Second, we now need extra constraints to ensure that the acquired impressions from a location are only assigned to campaigns which seek impressions from that location. Thus, as explained earlier in Section 5.2, we solve convex optimization problems to compute an upper bound on the true regret

Figure 4: Behavior of the regret with respect to the number of geographical locations ($L$) from where impressions are sought.

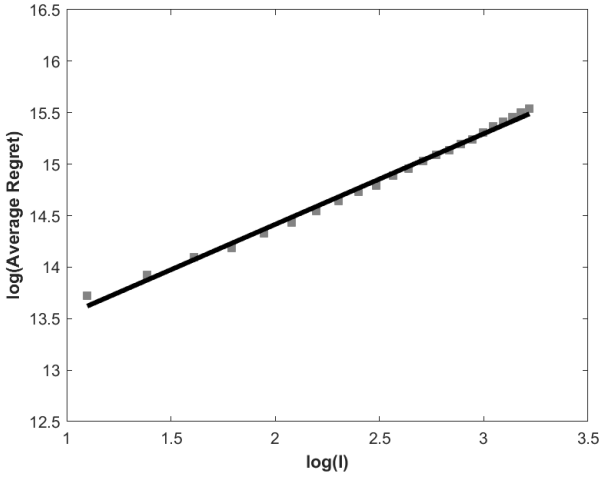



under our policy. In the generalized setting, the rate of increase in the regret with respect to $I$ is higher than that in the base case; specifically, the slope of the best-fit line in Figure 5 is about 0.88 as compared to 0.70 in Figure 2.

### 5.4 Decomposition of the Regret

In the numerical analysis reported in the previous subsection, we computed the regret as the difference between the expected cost under our policy and a *lower bound* on the optimal cost of the clairvoyant problem. Thus, our (reported) regret is, in fact, an upper bound on the "true regret" under our policy. A natural question arises: *How much of the regret is the true regret under our policy and how much of it is due to the use of a lower bound on the optimal cost of the clairvoyant problem (i.e., the gap between the optimal cost of the clairvoyant problem and its lower bound)*? Further, for the general setting where the unit penalty costs differ across campaigns, recall from Section 5.2 that we employ another approximation: Since the FEFS allocation policy is no longer optimal, we instead define and solve a convex optimization problem (based on the estimates of the parameters of the win curves) rather than solving the DP in the exploitation phase of our policy. The optimal solution of this problem is then used to obtain the bid price and allocation decision. This leads to another relevant question: *How much of the (true) regret is caused by learning (i.e., not knowing the parameters of the win curves) and how much of it is caused by the suboptimal allocation of impressions?* We examine both these questions numerically on a test bed of instances for the general setting where the unit penalty costs

Figure 5: Behavior of the regret with respect to the number of seasons, $I$, under the generalized setting




differ across campaigns.

For brevity, we relegate the details of this numerical study to Appendix M and only offer a quick summary of our results here. A substantial portion of the regret under our policy is due to the gap between the optimal cost of the clairvoyant problem and its lower bound. Specifically, after the first season, about 19% of the regret can be attributed to the use of the lower bound (instead of the optimal cost); after 50 seasons, this increases to about 36%. Our analysis of the second question shows that learning is the dominant cause of the regret. After the first season, about 85% of the regret is caused by learning and at most 15% of the regret is due to the suboptimal allocation of impressions. As time goes by, the learning of the win curves improves and we get progressively better estimates of the parameters of the win curves. Therefore, the percentage of the regret caused by learning reduces over time. After 50 seasons, about 66% of the regret is caused by learning.

## 6 Concluding Remarks

Our analysis in this paper is situated in the operations of a mobile-promotion platform that faces both supply-side and demand-side uncertainties. Each season, the platform accepts dynamically arriving campaigns from individual advertisers – a campaign requires the platform to deliver a certain number of mobile impressions from a set of locations over a desired time duration. The platform procures impressions via real-time bidding on an ad exchange. The platform learns the win curves at the various locations in real time based on the bids it places and the realized outcomes. Our two main results are: (1) An $\Omega(\sqrt{I})$ lower bound on the regret under *any* bidding and allocation policy, where $I$ is the number of seasons. (2) A bidding and allocation policy that offers a regret of $\mathcal{O}(\sqrt{I})$. Thus,

ours is an asymptotically tight learning algorithm for the platform.

In our setting, the mobile-promotion platform uses the bid prices for the impressions (at an ad-exchange) as a lever to learn unknown information on the supply side, namely the win curves at the various locations of interest. On the demand side, motivated by the current practice of fixed pricing of the campaigns, we assume that the campaigns arrive dynamically with a known probability distribution. However, to better match supply and demand, the platform can exploit pricing as a lever on the demand side. For example, the pricing of the campaigns may change dynamically depending on the length of the campaign, the number of required impressions, and the number of remaining periods in a season. The design of effective dynamic pricing schemes for the campaigns in conjunction with the learning of the win curves is an important and challenging direction in which future work can proceed.

## References


Agrawal N, Najafi Asadolahi S, Smith SA (2018) Optimization of Operational Decisions in Digital Advertising: A Literature Review. *Channel Strategies and Marketing Mix in a Connected World(Eds.),* Forthcoming.

Aseri M, Dawande M, Janakiraman G, Mookerjee V (2017) Procurement Policies for Mobile-Promotion Platforms. *Management Science* 64(10):4590–4607.

Baardman L, Fata E, Pani A, Perakis G (2019) Learning Optimal Online Advertising Portfolios with Periodic Budgets, working Paper.

Balseiro SR, Gur Y (2019) Learning in Repeated Auctions with Budgets: Regret Minimization and Equilibrium. *Management Science* 65(9):3952–3968.

Besbes O, Zeevi A (2012) Blind Network Revenue Management. *Operations Research* 60(6):1537–1550.

Borovkov A (1998) *Mathematical Statistics* (Amsterdam: Gordon and Breach).

Broder J, Rusmevichientong P (2012) Dynamic Pricing under a General Parametric Choice Model. *Operations Research* 60(4):965–980.

Chen B, Chao X, Ahn HS (2019a) Coordinating Pricing and Inventory Replenishment with Nonparametric Demand Learning. *Operations Research* 67(4):1035–1052.

Chen Q, Jasin S, Duenyas I (2019b) Nonparametric Self-Adjusting Control for Joint Learning and Optimization of Multiproduct Pricing with Finite Resource Capacity. *Mathematics of Operations Research* 44(2):601–631.

Chen YJ (2017) Optimal Dynamic Auctions for Display Advertising. *Operations Research* 65(4):897–913.

Choi H, Mela C, Balseiro S, Leary A (2019) Online Display Advertising Markets: A Literature Review and Future Directions. *Information Systems Research,* Forthcoming.

Cover TM, Thomas JA (2012) *Elements of Information Theory* (John Wiley & Sons).

den Boer AV (2015) Dynamic Pricing and Learning: Historical Origins, Current Research, and New Directions. *Surveys in Operations Research and Management Science* 20(1):1–18.

den Boer AV, Zwart B (2015) Dynamic Pricing and Learning with Finite Inventories. *Operations Research* 63(4):965–978.

eMarketer (2018) Mobile Ad Spend to Surpass All Traditional Media Combined by 2020. `https://www.emarketer.com/content/mobile-ad-spending-to-surpass-all-traditional-media-combined-by-2020`.

eMarketer (2019) Average US Time Spent with Mobile in 2019 Has Increased. https://www.emarketer.com/content/average-us-time-spent-with-mobile-in-2019-has-increased.

Harrison JM, Keskin NB, Zeevi A (2012) Bayesian Dynamic Pricing Policies: Learning and Earning under a Binary Prior Distribution. *Management Science* 58(3):570–586.

Harrison JM, Sunar N (2015) Investment Timing with Incomplete Information and Multiple Means of Learning. *Operations Research* 63(2):442–457.

Iyer K, Johari R, Sundararajan M (2014) Mean Field Equilibria of Dynamic Auctions with Learning. *Management Science* 60(12):2949–2970.

Keskin NB, Li M (2021) Selling Quality-Differentiated Products in a Markovian Market with Unknown Transition Probabilities. *Available at SSRN* .

Keskin NB, Li Y, Sunar N (2020) Data-driven Clustering and Feature-based Retail Electricity Pricing with Smart Meters. *Available at SSRN* .

Keskin NB, Zeevi A (2014) Dynamic Pricing with an Unknown Demand Model: Asymptotically Optimal Semi-Myopic Policies. *Operations Research* 62(5):1142–1167.

Korula N, Mirrokni V, Nazerzadeh H (2015) Optimizing Display Advertising Markets: Challenges and Directions. *IEEE Internet Computing* 20(1):28–35.

Levi R, Perakis G, Uichanco J (2015) The Data-Driven Newsvendor Problem: New Bounds and Insights. *Operations Research* 63(6):1294–1306.

Munkres JR (2018) *Analysis on Manifolds* (CRC Press).

Qi A, Ahn HS, Sinha A (2017) Capacity Investment with Demand Learning. *Operations Research* 65(1):145–164.

Sunar N, Yu S, Kulkarni VG (2021) Competitive Investment with Bayesian Learning: Choice of Business Size and Timing. *Operations Research* .

Tsybakov AB (2009) *Introduction to Nonparametric Estimation* (Springer).

Wang Z, Deng S, Ye Y (2014) Close the Gaps: A Learning-While-Doing Algorithm for Single-Product Revenue Management Problems. *Operations Research* 62(2):318–331.

Zhang W, Yuan S, Wang J (2014) Optimal Real-Time Bidding for Display Advertising. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1077–1086 (ACM).

## Online Appendix: Proofs and Additional Technical Results

## Appendix A History $h^{\pi}_{i,t} = h^{\pi}_{i,t}(x_{i,t})$ Corresponding to $x_{i,t}$ and Policy $\pi$ (Section 2)

For any $x_{i,t}$ (defined in (2)) and (deterministic) policy $\pi$, we show inductively that there is a unique corresponding $h^{\pi}_{i,t} = h^{\pi}_{i,t}(x_{i,t})$, where $h^{\pi}_{i,t}$ is defined in (1); see Section 2. In period $(1,1)$, for $x_{1,1} = (W_{1,j}, \bar{t}_{1,j}, \underline{t}_{1,j} : 1 \leq j \leq m_1)$, we have $h^{\pi}_{1,1}(x_{1,1}) = x_{1,1}$. Let $h^{\pi}_{i,t}(x_{i,t})$ be the history corresponding to $x_{i,t}$. Then, the bid price in period $(i,t)$ is $b_{i,t} = b^{\pi}_{i,t,l}(h^{\pi}_{i,t}(x_{i,t}))$ if $\zeta_{i,t} = l$ for $l \in \mathcal{L}$, and $b_{i,t} = 0$ if $\zeta_{i,t} = 0$. The allocation decision in period $(i,t)$ is $a_{i,t} = a^{\pi}_{i,t}(h^{\pi}_{i,t}(x_{i,t}))$. Thus, if $t < T$, then for $x_{i,t+1}$, we have $h^{\pi}_{i,t+1}(x_{i,t+1}) = (h^{\pi}_{i,t}(x_{i,t}), \zeta_{i,t}, b_{i,t}, d_{i,t}, a_{i,t})$. If $t = T$, then for $x_{i+1,1}$, we have

$$h^{\pi}_{i+1,1}(x_{i+1,1}) = (h^{\pi}_{i,T}(x_{i,T}), \zeta_{i,T}, b_{i,T}, d_{i,T}, a_{i,T}, W_{i+1,j}, \bar{t}_{i+1,j}, \underline{t}_{i+1,j} : 1 \leq j \leq m_{i+1}).$$

## Appendix B Obtaining the Number of Unmet Impressions $c_j(\hat{x}_{i,t})$ of Each Campaign $(i,j)$ at the Beginning of Period $(i,t)$, for a Given $\hat{x}_{i,t}$ (Section 2.3)

Recall from Section 2.3 that (a) $\hat{x}_{i,t}$ denotes the history in season $i$ until the beginning of period $(i,t)$ and (b) $c_j(\hat{x}_{i,t})$ denotes the number of unmet impressions of campaign $(i,j)$ in period $(i,t)$, where $t \leq T$. For any $\hat{x}_{i,t}$, we derive a recursive expression for $c_j(\hat{x}_{i,t})$. In period $(i,1)$, corresponding to $\hat{x}_{i,1} = \emptyset$, we have $c_j(\emptyset) = W_{i,j}$. Recall that $c(\hat{x}_{i,t})$ is the total number of unmet impressions at the beginning of period $(i,t)$ over all the ongoing campaigns in that period.

- If $c(\hat{x}_{i,t}) = 0$, then $c_j(\hat{x}_{i,t+1}) = c_j(\hat{x}_{i,t})$ for $j = 1, \cdots, m_i$.
- If $c(\hat{x}_{i,t}) > 0$, recall that the impression won in period $(i,t)$ is allocated to the active campaign that ends first, i.e., $(i, g_{i,t})$, where $g_{i,t} = \min_{(i,j) \in \mathcal{F}_{i,t}} j$. For any $l \in \mathcal{L}$ and $j \neq g_{i,t}$, $c_j(\hat{x}_{i,t+1}) = c_j(\hat{x}_{i,t})$. For any $l \in \mathcal{L}$, $c_{g_{i,t}}(\hat{x}_{i,t}, (0,0)) = c_{g_{i,t}}(\hat{x}_{i,t}, (l,0)) = c_{g_{i,t}}(\hat{x}_{i,t})$ and $c_{g_{i,t}}(\hat{x}_{i,t}, (l,1)) = c_{g_{i,t}}(\hat{x}_{i,t}) - 1$.

## Appendix C Proofs of Lemmas 1 Through 3

We first establish two auxiliary results, Lemmas A.1 and A.2, which will be used to prove Lemmas 1 through 3.

**Lemma A.1** *For any $l \in \mathcal{L}$, $b_l \in B$, $\alpha \in \mathbb{R}$ and $\gamma_l \in \Gamma_l$, define the function*

$$f^l(b_l, \alpha, \gamma_l) = p_l(\gamma_l, b_l)(b_l - e - \alpha).$$

*Let $b^*_l(\alpha, \gamma_l) = \arg\min_{b_l \in B} f^l(b_l, \alpha, \gamma_l)$. Thus, for $\mathcal{F} \neq \emptyset$, we have $b^*_{i,t,l}(\boldsymbol{c_i}; \gamma) = b^*_l(\alpha, \gamma_l)$, where $\alpha = \Delta V_{i,t+1}(\boldsymbol{c_i}; \gamma)$. Then:*

*(i) For each $(\alpha, \gamma_l) \in \mathbb{R} \times \Gamma^{(0)}_l$, $b^*_l(\alpha, \gamma_l)$ is uniquely defined.*

*(ii) Let* $\mathcal{U}_{A\Gamma_l} = \left\{(\alpha,\gamma_l) \in [b^{\min}-e,0] \times \Gamma_l^{(0)} \,\middle|\, b^{\min} < b_l^*(\alpha,\gamma_l) < b^{\max}\right\}$. *For each* $(\alpha,\gamma_l) \in \mathcal{U}_{A\Gamma_l}$,

$$\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)} > 0.$$

*(iii) For each* $(\alpha,\gamma_l) \in \mathcal{U}_{A\Gamma_l}$, *both* $b_l^*(\alpha,\gamma_l)$ *and* $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ *are continuously differentiable in* $\alpha$ *and* $\gamma_l$.

*(iv) For each* $(\alpha,\gamma_l) \in \mathcal{U}_{A\Gamma_l}$, $b_l^*(\alpha,\gamma_l)$ *is increasing in* $\alpha$.

*(v) There exists a* $K_0 > 0$ *such that* $f^l(b_l,\alpha,\gamma_l) - f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) \le K_0(b_l - b_l^*(\alpha,\gamma_l))^2$ *for all* $b_l \in B$ *and* $(\alpha,\gamma_l) \in \mathcal{U}_{A\Gamma_l}$.

**Proof of Lemma A.1**:

**(i)** Let $(\alpha,\gamma_l) \in \mathbb{R} \times \Gamma_l^{(0)}$. We have

$$\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l} = p_l(\gamma_l,b_l) + (b_l - e - \alpha)\frac{\partial p_l(\gamma_l,b_l)}{\partial b_l}, \tag{A-1}$$

and

$$\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2} = 2\frac{\partial p_l(\gamma_l,b_l)}{\partial b_l} + (b_l - e - \alpha)\frac{\partial^2 p_l(\gamma_l,b_l)}{\partial b_l^2}.$$

It follows that any $b_l \in B$ with $\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l} = 0$ satisfies the following:

$$\begin{aligned}
\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2} &= 2\frac{\partial p_l(\gamma_l,b_l)}{\partial b_l} + \frac{-p_l(\gamma_l,b_l)}{\partial p_l(\gamma_l,b_l)/\partial b_l}\frac{\partial^2 p_l(\gamma_l,b_l)}{\partial b_l^2} \\
&= \frac{\partial p_l(\gamma_l,b_l)}{\partial b_l}\left[2 - \frac{p_l(\gamma_l,b_l)\partial^2 p_l(\gamma_l,b_l)/\partial b_l^2}{(\partial p_l(\gamma_l,b_l)/\partial b_l)^2}\right] \\
&= \frac{\partial p_l(\gamma_l,b_l)}{\partial b_l}\left[1 + \frac{(\partial p_l(\gamma_l,b_l)/\partial b_l)^2 - p_l(\gamma_l,b_l)\partial^2 p_l(\gamma_l,b_l)/\partial b_l^2}{p_l(\gamma_l,b_l)^2}\frac{p_l(\gamma_l,b_l)^2}{(\partial p_l(\gamma_l,b_l)/\partial b_l)^2}\right] \\
&= \frac{\partial p_l(\gamma_l,b_l)}{\partial b_l}\left[1 - \frac{\partial^2 \log(p_l(\gamma_l,b_l))}{\partial b_l^2}\frac{p_l(\gamma_l,b_l)^2}{(\partial p_l(\gamma_l,b_l)/\partial b_l)^2}\right] \\
&> 0.
\end{aligned}$$

The inequality holds since $\frac{\partial p_l(\gamma_l,b_l)}{\partial b_l} > 0$ by Assumption 1 (Section 2.1) and $\frac{\partial^2 \log(p_l(\gamma_l,b_l))}{\partial b_l^2} \le 0$ by the log-concavity of $p_l(\gamma_l,b_l)$ with respect to $b_l$.

Thus, $f^l(b_l,\alpha,\gamma_l)$ either has a unique minimum $b_l^*(\alpha,\gamma_l) \in (b^{\min},b^{\max})$ with

$$\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)} > 0, \tag{A-2}$$

or is monotone on $B$ and the unique minimum of $f^l(b_l,\alpha,\gamma_l)$ is on the boundary of $B$.

**(ii)** For $(\alpha,\gamma_l) \in \mathcal{U}_{A\Gamma_l}$, since $b_l^*(\alpha,\gamma_l) \in (b^{\min},b^{\max})$, we have

$$\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)} > 0 \text{ by (A-2)}.$$

**(iii)** We first show that $b_l^*(\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$ using the Implicit Function Theorem (see, e.g., Theorem 9.2 in Munkres 2018). Notice that

- By Assumption 1, $\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}$ in Equation (A-1) is continuously differentiable in $\alpha$, $\gamma_l$, and $b_l$, on the open set $\mathbb{R}\times\Gamma_l\times(b^{\min},b^{\max})$ .

- By Lemma A.1 (ii), for each $(\alpha,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$, $(\alpha,\gamma_l,b_l^*(\alpha,\gamma_l))$ is a point in $\mathbb{R}\times\Gamma_l\times(b^{\min},b^{\max})$ such that

$$\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}=0 \text{ and } \left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)}>0.$$

Therefore, by the Implicit Function Theorem, $b_l^*(\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$.

Next, we show that $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$. The partial derivatives of $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ with respect to $\alpha$ and $\gamma_l$ are:

$$\begin{aligned}
\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial\alpha} &= \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}\frac{\partial b_l^*(\alpha,\gamma_l)}{\partial\alpha}+\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial\alpha}\right|_{b_l=b_l^*(\alpha,\gamma_l)}=-p_l(\gamma_l,b_l^*(\alpha,\gamma_l)),\\
\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial\gamma_l} &= \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}\frac{\partial b_l^*(\alpha,\gamma_l)}{\partial\gamma_l}+\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial\gamma_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}\\
&= (b_l^*(\alpha,\gamma_l)-e-\alpha)\left.\frac{\partial p_l(\gamma_l,b_l)}{\partial\gamma_l}\right|_{b_l^*(\alpha,\gamma_l)}.
\end{aligned}$$

By Assumption 1 and the fact that $b_l^*(\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$, the above expressions of $\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial\alpha}$ and $\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial\gamma_l}$ are continuous in $\alpha$ and $\gamma_l$. Thus, $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$.

**(iv)** For all $(\alpha,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$, we have

$$\frac{\partial b_l^*(\alpha,\gamma_l)}{\partial\alpha}=-\left(\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)}\right)^{-1}\cdot\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l\partial\alpha}\right|_{b_l=b_l^*(\alpha,\gamma_l)}.$$

Note that $\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)}>0$ by part (ii) of Lemma A.1 and

$$\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l\partial\alpha}\right|_{b_l=b_l^*(\alpha,\gamma_l)}=-\left.\frac{\partial p_l(\gamma_l,b_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}<0.$$

The strict inequality holds by Assumption 1. Thus, we have $\frac{\partial b_l^*(\alpha,\gamma_l)}{\partial\alpha}>0$, i.e., $b_l^*(\alpha,\gamma_l)$ is increasing in $\alpha$.

**(v)** Let $K_0^l:=\sup_{(\alpha,\gamma_l,b_l)\in\mathcal{U}_{A\Gamma_l}\times B}\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}/2$. Since $\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}$ is continuous in $\alpha$, $\gamma_l$, and $b_l$, on the closure of $\mathcal{U}_{A\Gamma_l}\times B$, which is compact, and $\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)}>0$ for all $(\alpha,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$, we have $0<K_0^l<\infty$. The Taylor expansion of $f^l(b_l,\alpha,\gamma_l)$ at $b_l=b_l^*(\alpha,\gamma_l)$ implies that

$$\begin{aligned}
f^l(b_l,\alpha,\gamma_l) &\le f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)+\left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)}(b_l-b_l^*(\alpha,\gamma_l))+K_0^l(b_l-b_l^*(\alpha,\gamma_l))^2\\
&= f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)+K_0^l(b_l-b_l^*(\alpha,\gamma_l))^2.
\end{aligned}$$

Let $K_0:=\max_{l\in\mathcal{L}}K_0^l$. Then, we have $f^l(b_l,\alpha,\gamma_l)-f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)\le K_0(b_l-b_l^*(\alpha,\gamma_l))^2$ for all $l\in\mathcal{L}$. ■

**Lemma A.2** *For $i\in\{1,\cdots,I\}$, $2\le t\le T+1$, $1\le j\le m_i$, and $\gamma\in\Gamma^{(0)}$, we have $b^{\min}-e\le V_{i,t}(\boldsymbol{c_i};\gamma)-V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma)\le 0$ where $\boldsymbol{e_j}$ is the unit vector of dimension $m_i$ whose $j^{th}$ component is 1.*

**Proof of Lemma A.2**: In the notation of Lemma A.1, let $\alpha = \Delta V_{i,t+1}(\boldsymbol{c_i};\gamma)$ and $\alpha' = \Delta V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma)$. For each location $l \in \mathcal{L}$, $b^*_{i,t,l}(\boldsymbol{c_i};\gamma) = b^*_l(\alpha,\gamma_l)$ and $b^*_{i,t,l}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) = b^*_l(\alpha',\gamma_l)$ are the optimal bids in period $(i,t)$ at states $\boldsymbol{c_i}$ and $\boldsymbol{c_i}-\boldsymbol{e_j}$, respectively.

The proof is by induction on $t$. For $t = T+1$, $V_{i,T+1}(\boldsymbol{c_i};\gamma) - V_{i,T+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) = 0$. For $2 \le t \le T$, suppose that $b^{\min} - e \le V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \le 0$. We now show that $b^{\min} - e \le V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \le 0$ using the following three cases. Let $\mathcal{F}_{i,t}(\boldsymbol{c_i})$ denote the set of all active campaigns in time period $(i,t)$ when $\boldsymbol{c_i}$ is the vector of the number of unmet impressions at the beginning of that period for each campaign in the season. Among the active campaigns in time period $(i,t)$, let $g_{i,t}(\boldsymbol{c_i})$ denote a campaign that ends first. For simplicity of exposition, we drop the indices $i$ and $t$ of $\mathcal{F}_{i,t}$ and $g_{i,t}$ below.

- Case 1: $\mathcal{F}(\boldsymbol{c_i}-\boldsymbol{e_j}) \neq \emptyset$. Then, $\mathcal{F}(\boldsymbol{c_i}) \neq \emptyset$. We first show that $V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \ge b^{\min} - e$:

$$
\begin{aligned}
& V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \\
= & \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l\in\mathcal{L}}} \left\{ \begin{array}{l} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)[b_l - e + V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_{g(c_i)}};\gamma)] + \\ [1-\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)] V_{i,t+1}(\boldsymbol{c_i};\gamma) \end{array} \right\} - \\
& \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l\in\mathcal{L}}} \left\{ \begin{array}{l} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)[b_l - e + V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j}-\boldsymbol{e_{g(c_i-e_j)}};\gamma)] + \\ [1-\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)] V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \end{array} \right\} \\
\ge & \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\alpha,\gamma_l))(V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_{g(c_i)}}) - V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j}-\boldsymbol{e_{g(c_i-e_j)}};\gamma)) + \\
& \left[1 - \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\alpha,\gamma_l))\right](V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma)) \\
\ge & \, b^{\min} - e.
\end{aligned}
$$

  The first inequality holds by letting $b_l = b^*_l(\alpha,\gamma_l)$ and the second holds by the induction hypothesis. Next, we show that $V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \le 0$:

$$
\begin{aligned}
& V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \\
= & \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l\in\mathcal{L}}} \left\{ \begin{array}{l} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)[b_l - e + V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_{g(c_i)}};\gamma)] + \\ [1-\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)] V_{i,t+1}(\boldsymbol{c_i};\gamma) \end{array} \right\} - \\
& \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l\in\mathcal{L}}} \left\{ \begin{array}{l} \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)[b_l - e + V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j}-\boldsymbol{e_{g(c_i-e_j)}};\gamma)] + \\ [1-\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)] V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma) \end{array} \right\} \\
\le & \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\alpha',\gamma_l))(V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_{g(c_i)}}) - V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j}-\boldsymbol{e_{g(c_i-e_j)}};\gamma)) + \\
& \left[1 - \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\alpha',\gamma_l))\right](V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}(\boldsymbol{c_i}-\boldsymbol{e_j};\gamma)) \\
\le & \, 0.
\end{aligned}
$$

The first inequality holds by letting $b_l = b_l^*(\alpha', \gamma_l)$ and the second holds by the induction hypothesis.

- Case 2: $\mathcal{F}(\boldsymbol{c_i} - \boldsymbol{e_j}) = \emptyset$ and $\mathcal{F}(\boldsymbol{c_i}) \neq \emptyset$. Then, we have $g(\boldsymbol{c_i}) = j$. We first show that $V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \geq b^{\min} - e$:

$$\begin{aligned}
& V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \\
= & \sum_{l \in \mathcal{L}} q_l \min_{b_l \in B_l} p_l(\gamma_l, b_l)[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i}; \gamma)] + V_{i,t+1}(\boldsymbol{c_i}; \gamma) - V_{i,t+1}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \\
= & \sum_{l \in \mathcal{L}} q_l p_l(\gamma_l, b_l^*(\alpha, \gamma_l))(b_l^*(\alpha, \gamma_l) - e) + \left[1 - \sum_{l \in \mathcal{L}} q_l p_l(\gamma_l, b_l^*(\alpha, \gamma_l))\right] \Delta V_{i,t+1}(\boldsymbol{c_i}) \\
\geq & \, b^{\min} - e.
\end{aligned}$$

  The inequality holds by the induction hypothesis.

  Next, we show that $V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \leq 0$:

$$\begin{aligned}
& V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \\
= & \sum_{l \in \mathcal{L}} q_l \min_{b_l \in B_l} p_l(\gamma_l, b_l)[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i}; \gamma)] + V_{i,t+1}(\boldsymbol{c_i}; \gamma) - V_{i,t+1}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \\
\leq & \sum_{l \in \mathcal{L}} q_l p_l(\gamma_l, b^{\min})[b^{\min} - e - \Delta V_{i,t+1}(\boldsymbol{c_i}; \gamma)] + V_{i,t+1}(\boldsymbol{c_i}; \gamma) - V_{i,t+1}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \\
\leq & \, 0.
\end{aligned}$$

  The last inequality holds by the induction hypothesis.

- Case 3: $\mathcal{F}(\boldsymbol{c_i} - \boldsymbol{e_j}) = \mathcal{F}(\boldsymbol{c_i}) = \emptyset$. In this case, we have $V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) = V_{i,t+1}(\boldsymbol{c_i}; \gamma) - V_{i,t+1}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma)$. Therefore, $b^{\min} - e \leq V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i} - \boldsymbol{e_j}; \gamma) \leq 0$ by the induction hypothesis. ■

**Proof of Lemma 1:** In the notation of Lemma A.1, let $\alpha = \Delta V_{i,t+1}(\boldsymbol{c_i}; \gamma)$ and let $\hat{\alpha} = \Delta V_{i,t+1}(\boldsymbol{c_i}; \hat{\gamma})$. Then, $b_{i,t,l}^*(\boldsymbol{c_i}; \gamma) - b_{i,t,l}^*(\boldsymbol{c_i}; \hat{\gamma}) = b_l^*(\alpha, \gamma_l) - b_l^*(\hat{\alpha}, \hat{\gamma}_l)$. By Lemma A.2 and Assumption 3, $b^{\min} - e \leq \alpha, \hat{\alpha} \leq 0$ and $b_l^*(\alpha, \gamma_l), b_l^*(\hat{\alpha}, \hat{\gamma}_l) \in (b^{\min}, b^{\max})$. Thus, $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$ and $(\hat{\alpha}, \hat{\gamma}) \in \mathcal{U}_{A\Gamma_l}$. Since $b_l^*(\alpha, \gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$ by part (iii) of Lemma A.1 and by the fact that the closure of $\mathcal{U}_{A\Gamma_l}$ is compact, it follows from the first-order Taylor expansion that

$$|b_l^*(\alpha, \gamma_l) - b_l^*(\hat{\alpha}, \hat{\gamma}_l)| \leq K_8^l(|\alpha - \hat{\alpha}| + \|\gamma_l - \hat{\gamma}_l\|), \tag{A-3}$$

for $K_8^l > 0$ that is independent of $\alpha$, $\hat{\alpha}$, $\gamma_l$, and $\hat{\gamma}_l$. We show by backward induction below (under the title "Derivation of Inequality (A-4)") that there exists $\kappa_t > 0$ such that

$$|V_{i,t}(\boldsymbol{c_i}; \gamma) - V_{i,t}(\boldsymbol{c_i}; \hat{\gamma})| \leq \kappa_t \|\gamma - \hat{\gamma}\|, \tag{A-4}$$

where $\kappa_t \leq (2K_9 + 1)^{T-1} - 1$ for all $t \geq 2$ and a positive constant $K_9$.

Combining (A-3) and (A-4), we have

$$
\begin{aligned}
&|b_l^*(\alpha,\gamma_l) - b_l^*(\hat{\alpha},\hat{\gamma}_l)| \\
&\leq K_8^l(|\alpha-\hat{\alpha}| + \|\gamma_l - \hat{\gamma}_l\|) \\
&\leq K_8^l(|V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}(\boldsymbol{c_i};\hat{\gamma})| + \\
&\quad |V_{i,t+1}((c_1,\ldots,c_g-1,\ldots,c_{m_i});\gamma) - V_{i,t+1}((c_1,\ldots,c_g-1,\ldots,c_{m_i});\hat{\gamma})| + \|\gamma_l - \hat{\gamma}_l\|) \\
&\leq K_8^l\left(2\left[(2K_9+1)^{T-1}-1\right]\|\gamma-\hat{\gamma}\| + \|\gamma-\hat{\gamma}\|\right) \\
&\leq (K_1)^T\|\gamma-\hat{\gamma}\|,
\end{aligned}
$$

where $K_1 = \max\left\{2\max_{l\in\mathcal{L}} K_8^l, 2K_9+1\right\}$. ■

**Derivation of Inequality** (A-4)**:** We show inequality (A-4) by backward induction on $t$. If $t = T+1$, then $V_{i,T+1}(\boldsymbol{c_i};\gamma) = V_{i,T+1}(\boldsymbol{c_i};\hat{\gamma}) = 0$ and (A-4) holds. Let $1 \leq t \leq T$. Suppose (A-4) holds for $t+1$. We now show that (A-4) holds for $t$.

$$
\begin{aligned}
&|V_{i,t}(\boldsymbol{c_i};\gamma) - V_{i,t}(\boldsymbol{c_i};\hat{\gamma})| \\
&= \left| \begin{array}{l} \mathbb{1}\{\mathcal{F}\neq\emptyset\}\sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\gamma_l,b_l)[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i};\gamma)] + V_{i,t+1}(\boldsymbol{c_i};\gamma) - \\ \mathbb{1}\{\mathcal{F}\neq\emptyset\}\sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\hat{\gamma}_l,b_l)[b_l - e - \Delta V_{i,t+1}(\boldsymbol{c_i};\hat{\gamma})] - V_{i,t+1}(\boldsymbol{c_i};\hat{\gamma}) \end{array} \right| \\
&\leq \left|\sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) - \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\hat{\alpha},\hat{\gamma}_l),\hat{\alpha},\hat{\gamma}_l)\right| + \kappa_{t+1}\|\gamma-\hat{\gamma}\| \\
&\leq K_9(|\alpha-\hat{\alpha}| + \|\gamma-\hat{\gamma}\|) + \kappa_{t+1}\|\gamma-\hat{\gamma}\| \\
&\leq K_9[|V_{i,t+1}(\boldsymbol{c_i};\gamma) - V_{i,t+1}(\boldsymbol{c_i};\hat{\gamma})| + \\
&\quad |V_{i,t+1}((c_1,\ldots,c_g-1,\ldots,c_{m_i});\gamma) - V_{i,t+1}((c_1,\ldots,c_g-1,\ldots,c_{m_i});\hat{\gamma})|] + (K_9+\kappa_{t+1})\|\gamma-\hat{\gamma}\| \\
&\leq K_9\left[\kappa_{t+1}\|\gamma-\hat{\gamma}\| + \kappa_{t+1}\|\gamma-\hat{\gamma}\|\right] + (K_9+\kappa_{t+1})\|\gamma-\hat{\gamma}\| \\
&= \kappa_t\|\gamma-\hat{\gamma}\|,
\end{aligned}
$$

where $\kappa_t = 2K_9\kappa_{t+1} + K_9 + \kappa_{t+1}$. The second inequality holds since $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$, by part (iii) of Lemma A.1. Therefore, $\sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma$. In addition, $[b^{\min}-e,0]\times\Gamma^{(0)}$ is compact. It follows by a first-order Taylor expansion that there exists $K_9 > 0$ that is independent of $\alpha$, $\hat{\alpha}$, $\gamma$, and $\hat{\gamma}$, such that

$$
\left|\sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) - \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\hat{\alpha},\hat{\gamma}_l),\hat{\alpha},\hat{\gamma}_l)\right| \leq K_9(|\alpha-\hat{\alpha}| + \|\gamma-\hat{\gamma}\|).
$$

Next, we show that $\kappa_t \leq (2K_9+1)^{T-t+1} - 1$ by backward induction on $t$. If $t = T+1$, then $\kappa_{T+1} = 0$. Let $1 \leq t \leq T$. Suppose $\kappa_{t+1} \leq (2K_9+1)^{T-t} - 1$. We now show that $\kappa_t \leq (2K_9+1)^{T-t+1} - 1$:

$$
\kappa_t = 2K_9\kappa_{t+1} + K_9 + \kappa_{t+1} \leq (2K_9+1)\left[(2K_9+1)^{T-t} - 1\right] + K_9 \leq (2K_9+1)^{T-t+1} - 1.
$$

Then, we have $\kappa_t \leq (2K_9+1)^{T-1}-1$ for all $t \geq 2$. ■

**Proof of Lemma 2:** We establish the result by showing that there exists a constant $K_0 > 0$ such that

$$V^{\pi}_{i,t}(\hat{x}_{i,t}; x_{i,1}, \gamma) - V_{i,t}(\boldsymbol{c_i}(\hat{x}_{i,t}); \gamma) \leq K_0 \mathbb{E}\left[ \sum_{\hat{t}=t}^{T} \sum_{l \in \mathcal{L}} q_l \left( b^{\pi}_{i,\hat{t},l}\left(x_{i,1}, \hat{X}^{\pi}_{i,\hat{t}}\right) - b^{*}_{i,\hat{t},l}\left(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}}); \gamma\right)\right)^2 \middle| \hat{X}^{\pi}_{i,t} = \hat{x}_{i,t} \right]. \tag{A-5}$$

Then, the regret under any FEFS policy $\pi$ after $I$ seasons satisfies

$$\begin{aligned}
&\sum_{i=1}^{I} \mathbb{E}\left[ V^{\pi}_{i,1}(\emptyset; X^{\pi}_{i,1}, \gamma^{(0)}) - V_{i,1}((W_{i,1}, \cdots, W_{i,m_i}); \gamma^{(0)}) \right] \\
=& \sum_{i=1}^{I} \mathbb{E}\left[ V^{\pi}_{i,1}(\emptyset; X^{\pi}_{i,1}, \gamma^{(0)}) - V_{i,1}(\boldsymbol{c_i}(\emptyset); \gamma^{(0)}) \right] \\
\leq& K_0 \mathbb{E}\left[ \sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} q_l \left( b^{\pi}_{i,t,l}\left(X^{\pi}_{i,1}, \hat{X}^{\pi}_{i,t}\right) - b^{*}_{i,t,l}\left(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,t}); \gamma^{(0)}\right)\right)^2 \right].
\end{aligned}$$

Next, we show (A-5) using backward induction on $t$. For $t = T+1$, we have

$$V^{\pi}_{i,T+1}(\hat{x}_{i,T+1}; x_{i,1}, \gamma) = V_{i,T+1}(\boldsymbol{c_i}(\hat{x}_{i,T+1}); \gamma) = 0.$$

Let $1 \leq t \leq T$. Suppose (A-5) holds at $t+1$, i.e.,

$$\begin{aligned}
&V^{\pi}_{i,t+1}(\hat{x}_{i,t+1}; x_{i,1}, \gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t+1}); \gamma) \\
\leq& K_0 \mathbb{E}\left[ \sum_{\hat{t}=t+1}^{T} \sum_{l \in \mathcal{L}} q_l \left( b^{\pi}_{i,\hat{t},l}\left(x_{i,1}, \hat{X}^{\pi}_{i,\hat{t}}\right) - b^{*}_{i,\hat{t},l}\left(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}}); \gamma\right)\right)^2 \middle| \hat{X}^{\pi}_{i,t+1} = \hat{x}_{i,t+1} \right].
\end{aligned}$$

Let $\alpha_l = V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (l,0)); \gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (l,1)); \gamma) = \Delta V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}); \gamma)$. Then, $b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t}); \gamma) = b^{*}_{l}(\alpha_l, \gamma_l)$. By Lemma A.2 and Assumption 3, we have $(\alpha_l, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$.

- Case 1: If $c(\hat{x}_{i,t}) = 0$, then we have

$$\begin{aligned}
&V^{\pi}_{i,t}(\hat{x}_{i,t}; x_{i,1}, \gamma) - V_{i,t}(\boldsymbol{c_i}(\hat{x}_{i,t}); \gamma) \\
=& \sum_{l \in \mathcal{L}} q_l V^{\pi}_{i,t+1}((\hat{x}_{i,t}, (l,0)); x_{i,1}, \gamma) + \left(1 - \sum_{l \in \mathcal{L}} q_l\right) V^{\pi}_{i,t+1}((\hat{x}_{i,t}, (0,0)); x_{i,1}, \gamma) - \\
& \sum_{l \in \mathcal{L}} q_l V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (l,0)); \gamma) - \left(1 - \sum_{l \in \mathcal{L}} q_l\right) V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (0,0)); \gamma) \\
=& \sum_{l \in \mathcal{L}} q_l \left[ V^{\pi}_{i,t+1}((\hat{x}_{i,t}, (l,0)); x_{i,1}, \gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (l,0)); \gamma) \right] + \\
& \left(1 - \sum_{l \in \mathcal{L}} q_l\right) \left[ V^{\pi}_{i,t+1}((\hat{x}_{i,t}, (0,0)); x_{i,1}, \gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t}, (0,0)); \gamma) \right] \\
=& \mathbb{E}\left[ V^{\pi}_{i,t+1}(\hat{X}^{\pi}_{i,t+1}; x_{i,1}, \gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,t+1}); \gamma) \middle| \hat{X}^{\pi}_{i,t} = \hat{x}_{i,t} \right] \\
\leq& \mathbb{E}\left[ K_0 \mathbb{E}\left[ \sum_{\hat{t}=t+1}^{T} \sum_{l \in \mathcal{L}} q_l \left( b^{\pi}_{i,\hat{t},l}(x_{i,1}, \hat{X}^{\pi}_{i,\hat{t}}) - b^{*}_{i,\hat{t},l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}}); \gamma)\right)^2 \middle| \hat{X}^{\pi}_{i,t+1} \right] \middle| \hat{X}^{\pi}_{i,t} = \hat{x}_{i,t} \right]
\end{aligned}$$

$$= K_0\mathbb{E}\left[\sum_{\hat{t}=t}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}}) - b^{*}_{i,\hat{t},l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}});\gamma)\right)^2 \middle| \hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right]$$

The inequality holds by the induction hypothesis.

- Case 2: If $c(\hat{x}_{i,t}) \geq 1$, then we have

$$V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma) - V_{i,t}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma)$$

$$= \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - e + V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma)\right] +$$
$$\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma) + \left(1-\sum_{l\in\mathcal{L}} q_l\right)V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma) -$$
$$\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma))\left[b^{*}_{i,t,l}(c(\hat{x}_{i,t});\gamma) - e + V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,1));\gamma)\right] -$$
$$\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma))\right]V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,0));\gamma) - \left(1-\sum_{l\in\mathcal{L}} q_l\right)V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(0,0));\gamma)$$

$$= \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,1));\gamma)\right] +$$
$$\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]\left[V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,0));\gamma)\right] +$$
$$\left(1-\sum_{l\in\mathcal{L}} q_l\right)\left[V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(0,0));\gamma)\right] +$$
$$\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,1));\gamma) + \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - e\right] +$$
$$\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,0));\gamma) -$$
$$\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma))\left[b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma) - e + V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,1));\gamma)\right] -$$
$$\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{x}_{i,t});\gamma))\right]V_{i,t+1}(\boldsymbol{c_i}(\hat{x}_{i,t},(l,0));\gamma)$$

$$= \mathbb{E}\left[V^{\pi}_{i,t+1}(\hat{X}^{\pi}_{i,t+1};x_{i,1},\gamma) - V_{i,t+1}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,t+1});\gamma)\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right] +$$
$$\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - e - \alpha_l\right] - \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^{*}_{l}(\alpha_l,\gamma_l))\left[b^{*}_{l}(\alpha_l,\gamma_l) - e - \alpha_l\right]$$

$$\leq \mathbb{E}\left[K_0\mathbb{E}\left[\sum_{\hat{t}=t+1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}}) - b^{*}_{i,\hat{t},l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}});\gamma)\right)^2\middle|\hat{X}^{\pi}_{i,t+1}\right]\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right] +$$
$$K_0\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - b^{*}_{l}(\alpha_l,\gamma_l)\right)^2$$

$$= K_0\mathbb{E}\left[\sum_{\hat{t}=t}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}}) - b^{*}_{i,\hat{t},l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,\hat{t}});\gamma)\right)^2\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right]$$

The inequality holds by the induction hypothesis and part (v) of Lemma A.1. ■

**Proof of Lemma 3**: For $l \in \mathcal{L}$, recall that $\hat{\gamma}_l(s)$ is the maximum-likelihood estimate of $\gamma_l$ based on the $sk_l$ observations corresponding to the placing of the exploration bids $\bar{\boldsymbol{b}}_{\boldsymbol{l}} = (\bar{b}_l^1, \cdots, \bar{b}_l^{k_l})$ repeatedly $s$ times. We apply the Tail Inequality (Theorem 36.3 in Borovkov 1998) on the finite-sample mean-squared error of maximum-likelihood estimators. For any $l \in \mathcal{L}$, Borovkov's result guarantees that there exist constants $\beta_1 > 0$ and $\beta_2 > 0$ such that for any $s \geq 1$ and any $\epsilon \geq 0$, we have

$$\Pr\left\{\left\|\hat{\gamma}(s) - \gamma^{(0)}\right\| \geq \epsilon\right\} \leq \beta_1 \exp(-s\beta_2\epsilon^2), \tag{A-6}$$

when the following conditions hold:

(i) The family $\{Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l} : \gamma_l \in \Gamma_l^{(0)}\}$ is identifiable.

(ii) For some $z > k_l$, $\sup_{\gamma_l \in \Gamma_l^{(0)}} \mathbb{E}\left[\left\|\nabla \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{D})\right\|^z\right] = \kappa < \infty$.

(iii) $\sqrt{Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d})}$ is differentiable in $\gamma_l$ on $\Gamma_l^{(0)}$ for any $\boldsymbol{d} \in \{0,1\}^{k_l}$.

(iv) The Fisher information matrix $\boldsymbol{I_l}(\bar{\boldsymbol{b}}_{\boldsymbol{l}}, \gamma_l)$ for any $\gamma_l \in \Gamma_l^{(0)}$ in (6) is positive definite.

The constants $\beta_1$ and $\beta_2$ depend only on $z$, $k_l$, $Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}$, and $\Gamma_l^{(0)}$.

To use inequality (A-6), we first verify that the above conditions hold. Conditions (i) and (iv) hold by Assumption 2. To verify condition (ii), recall that for any $\boldsymbol{d} \in \{0,1\}^{k_l}$,

$$Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d}) = \prod_{g=1}^{k_l} p_l(\gamma_l, \bar{b}_l^g)^{d_g}(1 - p_l(\gamma_l, \bar{b}_l^g))^{1-d_g}.$$

Thus, we have

$$\nabla \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d}) = \sum_{g=1}^{k_l} \left[d_g \nabla \log p_l(\gamma_l, \bar{b}_l^g) + (1 - d_g)\nabla \log(1 - p_l(\gamma_l, \bar{b}_l^g))\right],$$

which implies

$$\left\|\nabla \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d})\right\| \leq \sum_{g=1}^{k_l} \left[\left\|\nabla \log p_l(\gamma_l, \bar{b}_l^g)\right\| + \left\|\nabla \log(1 - p_l(\gamma_l, \bar{b}_l^g))\right\|\right].$$

Note that $p_l(\gamma_l, \bar{b}_l^g)$ is continuously differentiable in $\gamma_l$ on $\Gamma_l^{(0)}$ and is bounded away from 0 and 1 by Assumption 1. Thus, $\nabla \log p_l(\gamma_l, \bar{b}_l^g)$ and $\nabla \log(1 - p_l(\gamma_l, \bar{b}_l^g))$ are continuous in $\gamma_l$ on the compact set $\Gamma_l^{(0)}$. Consequently, there exists a constant $\bar{D}$ such that $\left\|\nabla \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d})\right\| \leq \bar{D}$. Then, with probability one, we have $\left\|\nabla \log Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d})\right\|^z \leq \bar{D}^z$. It follows that condition (ii) holds.

Next, we verify that condition (iii) holds. Note that $p_l(\gamma_l, \bar{b}_l^g)$ is differentiable in $\gamma_l$ on $\Gamma_l^{(0)}$ and is bounded away from 0 and 1 by Assumption 1. It follows that $Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d}) = \prod_{g=1}^{k_l} p_l(\gamma_l, \bar{b}_l^g)^{d_g}(1 - p_l(\gamma_l, \bar{b}_l^g))^{1-d_g}$ is also differentiable in $\gamma_l$ on $\Gamma_l^{(0)}$ and is bounded away from zero. Thus, $\sqrt{Q_l^{\bar{\boldsymbol{b}}_{\boldsymbol{l}},\gamma_l}(\boldsymbol{d})}$ is differentiable in $\gamma_l$ on $\Gamma_l^{(0)}$, i.e., condition (iii) holds.

Having verified conditions (i)-(iv) above, we now use (A-6) to show that there exists a constant $K^l_{mle} = \beta_1/\beta_2 > 0$ such that $\mathbb{E}\left[\left\|\hat{\gamma}_l(s) - \gamma_l^{(0)}\right\|^2\right] \leq \frac{K^l_{mle}}{s}$:

$$\mathbb{E}\left[\left\|\hat{\gamma}_l(s) - \gamma_l^{(0)}\right\|^2\right] = \int_0^\infty \Pr\left\{\left\|\hat{\gamma}(s) - \gamma^{(0)}\right\|^2 \geq \mu\right\} d\mu \leq \int_0^\infty \beta_1 \exp(-s\beta_2\mu) d\mu = \frac{\beta_1}{\beta_2 s} = \frac{K^l_{mle}}{s}.$$

Letting $K_{mle} = \sum_{l\in\mathcal{L}} K^l_{mle}$, we have

$$\mathbb{E}\left[\left\|\hat{\gamma}(s) - \gamma^{(0)}\right\|^2\right] = \sum_{l\in\mathcal{L}} \mathbb{E}\left[\left\|\hat{\gamma}_l(s) - \gamma_l^{(0)}\right\|^2\right] \leq \sum_{l\in\mathcal{L}} \frac{K^l_{mle}}{s} = \frac{K_{mle}}{s}. \qquad \blacksquare$$

## Appendix D Proofs of Theorems 2 and 3

Let $\gamma_0 := 1/2$; this constant plays an important role in the proofs of Theorems 2 and 3. Before we proceed to establish Theorem 2 (resp., Theorem 3), we state and prove an intermediate result, namely Lemma A.3 (resp., Lemma A.4), using the KL divergence as a measure of the difference between two distributions. Broder and Rusmevichientong (2012) also apply KL divergence to establish Theorem 3.1 in their paper. However, they compute the KL divergence of the distributions of the demands (consumer responses to a sequence of prices) under two different values of the underlying parameters, while we use the KL divergence of the joint distributions of the outcomes of impression arrivals (uncertain but observable) and the winning of impressions (responses to a sequence of bids) under two different values of $\gamma^{(0)}$.

The following definition is reproduced verbatim from Broder and Rusmevichientong (2012) (Definition 3.2 of their paper), who attribute the definition in this form to Cover and Thomas (1999).

**Definition 1 (Definition 2.26 in Cover and Thomas 1999).** *For any probability measures $Q_0$ and $Q_1$ on a discrete sample space $\mathcal{Y}$, the KL divergence of $Q_0$ and $Q_1$ is*

$$\mathcal{K}(Q_0; Q_1) = \sum_{y\in\mathcal{Y}} Q_0(y) \log\left(\frac{Q_0(y)}{Q_1(y)}\right).$$

Intuitively, if the KL divergence between two distributions is large, then they are far apart and easy to distinguish, and vice versa.

### D.1 Proof of Theorem 2

Consider the problem instance defined in the statement of Theorem 2. We compute the joint probability distribution of the realizations of impression arrival and winning outcomes under a given parameter $\gamma \in \Gamma^{(0)}$ and policy $\pi$. We then compute the KL divergence of the joint probability distributions corresponding to two different underlying parameters.

Recall that $T = 1$ in our instance; i.e., there is one period in each season. Let $\zeta_i = 1$ if an impression arrives in season $i$ and $\zeta_i = 0$ otherwise. Let $d_i = 1$ if an impression is won in season $i$ and $d_i = 0$ otherwise.

Let $x_i = (\zeta_i, d_i)$ denote the outcome in season $i$ and $\mathbb{X} = \{(0,0),(1,0),(1,1)\}$ denote the set of all possible outcomes in each season. Let $\boldsymbol{x_i} = (x_{\hat{i}} : 1 \leq \hat{i} \leq i)$ denote the outcome in the first $i$ seasons. Then, the bid price in season $i$ depends on the outcome in the past $i-1$ seasons, denoted by $b_i^\pi(\boldsymbol{x_{i-1}})$. For any $\gamma \in \Gamma^{(0)}$ and policy $\pi$, let

$$Q_i^{\pi,\gamma}(\boldsymbol{x_i}) = \prod_{\hat{i}=1}^{i} \left( \left[ qp(b_{\hat{i}}^\pi(\boldsymbol{x_{\hat{i}-1}}),\gamma)^{d_{\hat{i}}}(1-p(b_{\hat{i}}^\pi(\boldsymbol{x_{\hat{i}-1}}),\gamma))^{1-d_{\hat{i}}} \right]^{\zeta_{\hat{i}}} (1-q)^{1-\zeta_{\hat{i}}} \right).$$

Note that this is the probability of observing the realization $\boldsymbol{x_i}$ under policy $\pi$ when the underlying parameter is $\gamma$. Then, for any $\gamma \in \Gamma^{(0)}$, the KL divergence of $Q_i^{\pi,\gamma_0}$ and $Q_i^{\pi,\gamma}$ is

$$\mathcal{K}(Q_i^{\pi,\gamma_0}; Q_i^{\pi,\gamma}) = \sum_{\boldsymbol{x_i} \in \mathbb{X}^i} Q_i^{\pi,\gamma_0}(\boldsymbol{x_i}) \log \left( \frac{Q_i^{\pi,\gamma_0}(\boldsymbol{x_i})}{Q_i^{\pi,\gamma}(\boldsymbol{x_i})} \right).$$

The following result helps us make a connection between the regret of a policy $\pi$ and the above KL divergence.

**Lemma A.3** *Let $\gamma_1 = \gamma_0 + \frac{1}{4} I^{-1/4}$. For any $I \geq 2$ and policy $\pi$, the following statements hold:*

*(i) Regret$(\pi, I; \gamma_0) \geq \frac{7}{12}\sqrt{I}\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1})$.*

*(ii) Regret$(\pi, I; \gamma_0)$ + Regret$(\pi, I; \gamma_1) \geq q\frac{\sqrt{I}}{12(24^2)} \exp(-\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}))$.*

The proof of Lemma A.3 is provided in Appendix E. We now prove Theorem 2 using Lemma A.3.

**Proof of Theorem 2:** (i) It is easy to verify that Assumptions 1 and 2 are satisfied for the instance defined in the statement of the theorem. Next, we show that Assumption 3 is satisfied as well. Since $T = 1$ for the instance, we have

$$b_{i,1,l}^*(\boldsymbol{c_i}; \gamma) = \arg\min_{b \in B} p(b,\gamma)(b-e) = \arg\min_{b \in B} (1/2 - \gamma + \gamma b)(b-2) = 3/2 - 1/(4\gamma).$$

Let $b^*(\gamma) = 3/2 - 1/(4\gamma)$. Thus, $b^*(\gamma) \in [3/4, 5/4]$ for $\gamma \in [1/3, 1]$, which implies that $b^*(\gamma) \in (b^{\min}, b^{\max})$, i.e., Assumption 3 is satisfied.

(ii) Using part (i) of Lemma A.3, we have

$$\text{Regret}(\pi, I; \gamma_0) + \text{Regret}(\pi, I; \gamma_1) \geq \text{ Regret}(\pi, I; \gamma_0) \geq \frac{7}{12}\sqrt{I}\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}).$$

Combining this inequality and part (ii) of Lemma A.3, we have

$$\begin{aligned}
& 2[\text{Regret}(\pi, I; \gamma_0) + \text{Regret}(\pi, I; \gamma_1)] \\
\geq\ & \frac{7}{12}\sqrt{I}\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}) + q\frac{\sqrt{I}}{12(24^2)} \exp(-\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1})) \\
\geq\ & q\frac{\sqrt{I}}{12(24^2)} \left( \mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}) + \exp(-\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1})) \right) \\
\geq\ & q\frac{\sqrt{I}}{12(24^2)}. \qquad \text{(A-7)}
\end{aligned}$$

The last inequality holds since $\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}) \geq 0$ and $\chi + \exp(-\chi) \geq 1$ for any $\chi \geq 0$.

Inequality (A-7) implies that at least one of $\text{Regret}(\pi, I; \gamma_0)$ and $\text{Regret}(\pi, I; \gamma_1)$ is no less than $q\frac{\sqrt{I}}{2(24^3)}$. Therefore, there exists $\gamma^{(0)} \in \{\gamma_0, \gamma_1\}$ such that $\text{Regret}(\pi, I; \gamma^{(0)}) \geq q\frac{\sqrt{I}}{2(24^3)}$. ■

## D.2 Proof of Theorem 3

Consider the problem instance defined in the statement of Theorem 3. Note that $T = 2$ in this instance; i.e., there are two periods in each season. Recall that $\zeta_{i,t} = 1$ if an impression arrives in period $t \in \{1, 2\}$ of season $i$ and $\zeta_{i,t} = 0$ otherwise; $d_{i,t} = 1$ if an impression is won in period $t$ of season $i$ and $d_{i,t} = 0$ otherwise. Let $x_{i,t} = (\zeta_{i,t}, d_{i,t})$ denote the outcome in period $(i, t)$ and $\mathbb{X} = \{(0, 0), (1, 0), (1, 1)\}$ denote the set of all possible outcomes in a period. Let $\boldsymbol{x_{i,t}} = (x_{1,1}, x_{1,2}, \cdots, x_{i,t})$ denote the outcomes until period $t$ of season $i$. Then, the bid price in period 1 (resp., period 2) of season $i$, denoted by $b_{i,1}^{\pi}(\boldsymbol{x_{i-1,2}})$ (resp., $b_{i,2}^{\pi}(\boldsymbol{x_{i,1}})$), depends on the outcome in the past $i - 1$ seasons (resp., plus the outcome in period 1 of season $i$). For any $\gamma \in \Gamma^{(0)}$ and policy $\pi$, let

$$Q_{i,1}^{\pi,\gamma}(\boldsymbol{x_{i,1}}) = P_{i,1}^{\pi,\gamma}(\boldsymbol{x_{i,1}}) \prod_{\hat{i}=1}^{i-1} \left[ P_{\hat{i},1}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},1}}) P_{\hat{i},2}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},2}}) \right],$$

$$Q_{i,2}^{\pi,\gamma}(\boldsymbol{x_{i,2}}) = \prod_{\hat{i}=1}^{i} \left[ P_{\hat{i},1}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},1}}) P_{\hat{i},2}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},2}}) \right],$$

where

$$P_{\hat{i},1}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},1}}) = \left[ qp(b_{\hat{i},1}^{\pi}(\boldsymbol{x_{\hat{i}-1,2}}), \gamma)^{d_{\hat{i},1}} (1 - p(b_{\hat{i},1}^{\pi}(\boldsymbol{x_{\hat{i}-1,2}}), \gamma))^{1-d_{\hat{i},1}} \right]^{\zeta_{\hat{i},1}} (1 - q)^{1-\zeta_{\hat{i},1}},$$

$$P_{\hat{i},2}^{\pi,\gamma}(\boldsymbol{x_{\hat{i},2}}) = \left[ qp(b_{\hat{i},2}^{\pi}(\boldsymbol{x_{\hat{i},1}}), \gamma)^{d_{\hat{i},2}} (1 - p(b_{\hat{i},2}^{\pi}(\boldsymbol{x_{\hat{i},1}}), \gamma))^{1-d_{\hat{i},2}} \right]^{\zeta_{\hat{i},2}} (1 - q)^{1-\zeta_{\hat{i},2}}.$$

Then, for any $\gamma \in \Gamma^{(0)}$, the KL divergence of $Q_{i,t}^{\pi,\gamma_0}$ and $Q_{i,t}^{\pi,\gamma}$ is

$$\mathcal{K}(Q_{i,t}^{\pi,\gamma_0}; Q_{i,t}^{\pi,\gamma}) = \sum_{\boldsymbol{x_{i,t}} \in \mathbb{X}^{2i-2+t}} Q_{i,t}^{\pi,\gamma_0}(\boldsymbol{x_{i,t}}) \log \left( \frac{Q_{i,t}^{\pi,\gamma_0}(\boldsymbol{x_{i,t}})}{Q_{i,t}^{\pi,\gamma}(\boldsymbol{x_{i,t}})} \right).$$

The following result helps us make a connection between the regret of a policy $\pi$ and the above KL divergence. Recall that $\gamma_0 = 1/2$.

**Lemma A.4** *Let $\gamma_1 = \gamma_0 + \frac{1}{4} I^{-2/7}$. For any $I \geq 2$ and policy $\pi$, the following statements hold:*

*(i)* $\sqrt{\text{Regret}(\pi, I; \gamma_0)} \geq \frac{7I^{1/7}}{62\sqrt{K_4}} \mathcal{K}(Q_{I,2}^{\pi,\gamma_0}; Q_{I,2}^{\pi,\gamma_1}) - \frac{(K_4)^{5/2} I^{1/7}}{124}$.

*(ii)* $\text{Regret}(\pi, I; \gamma_0) + \text{Regret}(\pi, I; \gamma_1) \geq \frac{K_4 I^{2/7}}{4(24^3)} \exp(-\mathcal{K}(Q_{I,2}^{\pi,\gamma_0}; Q_{I,2}^{\pi,\gamma_1}))$.

The proof of Lemma A.4 is provided in Appendix F. We now establish Theorem 3 using Lemma A.4.

**Proof of Theorem 3:** (i) It is easy to verify that Assumptions 1 and 2 are satisfied for the instance defined in the statement of the theorem. Next, we show that Assumption 3 is satisfied as well:

$$b^*_{i,1,l}(1;\gamma) = \arg\min_{b\in B}(1/2-\gamma+\gamma b)(b-2-\Delta V_{i,2}(1;\gamma)) = \frac{3\gamma-1/2-q\gamma^2/4-q/16-q\gamma/4}{2\gamma},$$

$$b^*_{i,2,l}(1;\gamma) = \arg\min_{b\in B} p(b,\gamma)(b-e) = \arg\min_{b\in B}(1/2-\gamma+\gamma b)(b-2) = 3/2-1/(4\gamma).$$

Let $b^*_1(\gamma) = \frac{3\gamma-1/2-q\gamma^2/4-q/16-q\gamma/4}{2\gamma}$ and $b^*_2(\gamma) = 3/2-1/(4\gamma)$. It is easy to verify that $b^*_1(\gamma), b^*_2(\gamma) \in (b^{\min}, b^{\max})$ for $\gamma \in [1/3, 1]$, i.e., Assumption 3 is satisfied.

(ii) Using part (i) of Lemma A.4, we have

$$\sqrt{\text{Regret}(\pi, I;\gamma_0)+\text{Regret}(\pi, I;\gamma_1)} \geq \sqrt{\text{Regret}(\pi, I;\gamma_0)} \geq \frac{7I^{1/7}}{62\sqrt{K_4}}\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2}) - \frac{(K_4)^{5/2}I^{1/7}}{124}.$$

Combining this inequality and part (ii) of Lemma A.4, we have

$$\begin{aligned}
&2\sqrt{\text{Regret}(\pi, I;\gamma_0)+\text{Regret}(\pi, I;\gamma_1)} \\
&\geq \frac{7}{62\sqrt{K_4}} I^{1/7}\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2}) + \sqrt{K_4}\frac{I^{1/7}}{96\sqrt{6}}\exp(-\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2})/2) - \frac{(K_4)^{5/2}}{124}I^{1/7} \\
&\geq \sqrt{K_4}\frac{I^{1/7}}{96\sqrt{6}}\left(\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2}) + \exp(-\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2})/2)\right) - \frac{(K_4)^{5/2}}{124}I^{1/7} \\
&\geq \left[\frac{\sqrt{K_4}}{96\sqrt{6}} - \frac{(K_4)^{5/2}}{124}\right] I^{1/7}.
\end{aligned}$$

The last inequality holds since $\mathcal{K}(Q^{\pi,\gamma_0}_{I,2}; Q^{\pi,\gamma_1}_{I,2}) \geq 0$ and $\chi + \exp(-\chi/2) \geq 1$ for any $\chi \geq 0$. Note that $\frac{\sqrt{K_4}}{96\sqrt{6}} - \frac{(K_4)^{5/2}}{124} > 0$ for $0 < K_4 < \sqrt{\frac{31}{24\sqrt{6}}}$. Therefore, at least one of $\text{Regret}(\pi, I;\gamma_0)$ and $\text{Regret}(\pi, I;\gamma_1)$ is no less than

$$\frac{1}{8}\left[\frac{\sqrt{K_4}}{96\sqrt{6}} - \frac{(K_4)^{5/2}}{124}\right]^2 I^{2/7}. \qquad \blacksquare$$

## Appendix E Proof of Lemma A.3

We first derive some preliminary results that will be useful in establishing Lemma A.3.

Consider the instance of problem $(P)$ defined in Theorem 2. Also, recall from Section 3.3 that $\gamma_0 = 1/2$. We make the following observations:

- $p(\gamma, b) \in [1/8, 7/8]$ for any $b \in B$ and $\gamma \in \Gamma^{(0)}$ and hence

$$p(\gamma, b)(1-p(\gamma, b)) \geq 7/64. \tag{A-8}$$

- Let $b^\pi_i$ denote the bid price in season $i$ under policy $\pi$. Since the target of each campaign can never be exceeded, we can write the regret as:

$$\text{Regret}(\pi, I;\gamma^{(0)})$$

$$= \sum_{i=1}^{I} \mathbb{E}\big[q(b_i^\pi - e)p(\gamma^{(0)}, b_i^\pi)\big] - Iq \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\}$$
$$= q\sum_{i=1}^{I} \mathbb{E}\big[r(b_i^\pi, \gamma^{(0)}) - r(b^*(\gamma^{(0)}), \gamma^{(0)})\big], \tag{A-9}$$

where $r(b,\gamma) = (b-e)p(\gamma,b) = (b-2)(1/2 - \gamma + \gamma b)$ and $b^*(\gamma) = 3/2 - 1/(4\gamma)$ is the optimal solution that minimizes $r(b,\gamma)$.

- $b^*(\gamma_0) = 1$.
- Since $\gamma \le 1$, we have

$$|b^*(\gamma) - b^*(\gamma_0)| = \frac{|\gamma - \gamma_0|}{4\gamma\gamma_0} \ge \frac{|\gamma - \gamma_0|}{2}. \tag{A-10}$$

- The absolute difference $|p(\gamma_0, b) - p(\gamma, b)|$ satisfies:

$$|p(\gamma_0, b) - p(\gamma, b)| = |\gamma - \gamma_0||b-1| = |\gamma - \gamma_0||b - b^*(\gamma_0)|. \tag{A-11}$$

- $\frac{\partial^2 r(b,\gamma)}{\partial b^2} = 2\gamma \ge 2/3$. Using this along with the fact that $\frac{\partial r(b,\gamma)}{\partial b} = 0$ at $b = b^*(\gamma)$, we obtain

$$\begin{aligned} r(b,\gamma) - r(b^*(\gamma),\gamma) &= \left| \int_{b^*(\gamma)}^{b} \int_{b^*(\gamma)}^{v} \frac{\partial^2 r(\tilde{b},\gamma)}{\partial \tilde{b}^2} d\tilde{b} dv \right| \\ &\ge \frac{2}{3} \left| \int_{b^*(\gamma)}^{b} \int_{b^*(\gamma)}^{v} d\tilde{b} dv \right| \\ &= \frac{1}{3}(b - b^*(\gamma))^2. \end{aligned} \tag{A-12}$$

Lemma A.5 below proves an inequality that will be useful in establishing Lemma A.6. In turn, Lemma A.6 will be used to prove part (i) of Lemma A.3.

**Lemma A.5** *For any $\omega, \nu > 0$, $\omega \log(\frac{\omega}{\nu}) + (1-\omega)\log(\frac{1-\omega}{1-\nu}) \le \frac{(\omega-\nu)^2}{\nu(1-\nu)}$.*

**Proof of Lemma A.5:** Note that for any $\rho > 0$, $\log\rho \le \rho - 1$. Thus, we have $\omega \log(\frac{\omega}{\nu}) + (1-\omega)\log(\frac{1-\omega}{1-\nu}) \le \omega(\frac{\omega}{\nu} - 1) + (1-\omega)(\frac{1-\omega}{1-\nu} - 1) = \frac{(\omega-\nu)^2}{\nu(1-\nu)}$. ■

**Lemma A.6** *For any $\gamma \in \Gamma^{(0)}$, $i \ge 1$, and policy $\pi$,*

$$\sum_{\boldsymbol{x_{i-1}} \in \mathbb{X}^{i-1}} Q_{i-1}^{\pi,\gamma_0}(\boldsymbol{x_{i-1}}) \sum_{x_i \in \mathbb{X}} Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}}) \log\left(\frac{Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}})}{Q_i^{\pi,\gamma}(x_i|\boldsymbol{x_{i-1}})}\right) \le \frac{192}{7} q(\gamma_0 - \gamma)^2 \mathbb{E}\left[r(b_i^\pi, \gamma_0) - r(b^*(\gamma_0), \gamma_0)\right].$$

**Proof of Lemma A.6:** We first derive an upper bound on $\sum_{x_i \in \mathbb{X}} Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}}) \log\left(\frac{Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}})}{Q_i^{\pi,\gamma}(x_i|\boldsymbol{x_{i-1}})}\right)$:

$$\begin{aligned} &\sum_{x_i \in \mathbb{X}} Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}}) \log\left(\frac{Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}})}{Q_i^{\pi,\gamma}(x_i|\boldsymbol{x_{i-1}})}\right) \\ &= qp(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1})) \log\left(\frac{p(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1}))}{p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1}))}\right) + q(1 - p(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1}))) \log\left(\frac{1 - p(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1}))}{1 - p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1}))}\right) \end{aligned}$$

$$
\begin{aligned}
&\leq q\frac{(p(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1})) - p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1})))^2}{p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1}))(1-p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1})))} \\
&\leq \frac{64}{7}q(p(\gamma_0, b_i^\pi(\boldsymbol{x}_{i-1})) - p(\gamma, b_i^\pi(\boldsymbol{x}_{i-1})))^2 \\
&= \frac{64}{7}q(\gamma_0-\gamma)^2(b_i^\pi(\boldsymbol{x_{i-1}}) - b^*(\gamma_0))^2 \\
&\leq \frac{192}{7}q(\gamma_0-\gamma)^2(r(b_i^\pi(\boldsymbol{x_{i-1}}), \gamma_0) - r(b^*(\gamma_0), \gamma_0)). \qquad \text{(A-13)}
\end{aligned}
$$

The first inequality holds by Lemma A.5. The second inequality holds by (A-8). The second equality holds by (A-11). The last inequality holds by (A-12).

Using inequality (A-13), we have

$$
\begin{aligned}
&\sum_{\boldsymbol{x_{i-1}}\in\mathbb{X}^{i-1}} Q_{i-1}^{\pi,\gamma_0}(\boldsymbol{x_{i-1}}) \sum_{x_i\in\mathbb{X}} Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}}) \log\left(\frac{Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}})}{Q_i^{\pi,\gamma}(x_i|\boldsymbol{x_{i-1}})}\right) \\
&\leq \frac{192}{7}q(\gamma_0-\gamma)^2 \sum_{\boldsymbol{x_{i-1}}\in\mathbb{X}^{i-1}} Q_{i-1}^{\pi,\gamma_0}(\boldsymbol{x_{i-1}})(r(b_i^\pi(\boldsymbol{x_{i-1}}), \gamma_0) - r(b^*(\gamma_0), \gamma_0)) \\
&= \frac{192}{7}q(\gamma_0-\gamma)^2\mathbb{E}\left[r(b_i^\pi, \gamma_0) - r(b^*(\gamma_0), \gamma_0)\right]. \qquad \blacksquare
\end{aligned}
$$

The proof of part (ii) of Lemma A.3 uses the following result, which we reproduce verbatim from Lemma EC.1.3 in Broder and Rusmevichientong (2012). They obtain the lemma using Theorem 2.2 of Tsybakov (2009).

**Lemma A.7 (Theorem 2.2, Tsybakov 2009)** *Let $Q_0$ and $Q_1$ be two probability distributions on a finite space $\mathcal{Y}$, with $Q_0(y), Q_1(y) > 0$ for all $y \in \mathcal{Y}$. Then for any function $J : \mathcal{Y} \to \{0, 1\}$,*

$$
Q_0\{J = 1\} + Q_1\{J = 0\} \geq \frac{1}{2}\exp(-\mathcal{K}(Q_0; Q_1)),
$$

*where $\mathcal{K}(Q_0; Q_1)$ denotes the KL divergence of $Q_0$ and $Q_1$.*

**Proof of Lemma A.3:** (i) We apply the Chain rule for KL divergence (Theorem 2.5.3 in Cover and Thomas 2012):

$$
\begin{aligned}
\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma}) &= \sum_{i=1}^{I} \sum_{\boldsymbol{x_{i-1}}\in\mathbb{X}^{i-1}} Q_{i-1}^{\pi,\gamma_0}(\boldsymbol{x_{i-1}}) \sum_{x_i\in\mathbb{X}} Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}}) \log\left(\frac{Q_i^{\pi,\gamma_0}(x_i|\boldsymbol{x_{i-1}})}{Q_i^{\pi,\gamma}(x_i|\boldsymbol{x_{i-1}})}\right) \\
&\leq \frac{192}{7}(\gamma_0-\gamma)^2 q \sum_{i=1}^{I} \mathbb{E}[r(b_i^\pi, \gamma_0) - r(b^*(\gamma_0), \gamma_0)].
\end{aligned}
$$

The inequality holds by Lemma A.6. Thus, for $\gamma_1 = \gamma_0 + \frac{1}{4}I^{-1/4}$, using (A-9), we have

$$
\text{Regret}(\pi, I; \gamma_0) = q\sum_{i=1}^{I}\mathbb{E}\big[r(b_i^\pi, \gamma_0) - r(b^*(\gamma_0), \gamma_0)\big] \geq \frac{7}{192(\gamma_0-\gamma_1)^2}\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}) = \frac{7}{12}\sqrt{I}\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1}).
$$

(ii) We first define two intervals $B_{\gamma_0} \in B$ and $B_{\gamma_1} \in B$:

$$
B_{\gamma_0} := \left\{b : |b^*(\gamma_0) - b| \leq \frac{1}{24I^{1/4}}\right\} \text{ and } B_{\gamma_1} := \left\{b : |b^*(\gamma_1) - b| \leq \frac{1}{24I^{1/4}}\right\}.
$$

Note that $B_{\gamma_0}$ and $B_{\gamma_1}$ are disjoint since $|b^*(\gamma_1) - b^*(\gamma_0)| \geq \frac{|\gamma_1 - \gamma_0|}{2} = \frac{1}{8I^{1/4}}$ using (A-10). Recall from (A-12) that $r(b^*(\gamma), \gamma) - r(b, \gamma) \geq \frac{1}{3}(b^*(\gamma) - b)^2$. For each $\gamma \in \{\gamma_0, \gamma_1\}$, if $b \notin B_\gamma$, then

$$r(b^*(\gamma), \gamma) - r(b, \gamma) \geq \frac{1}{3}(b^*(\gamma) - b)^2 \geq \frac{1}{3(24^2)\sqrt{I}}.$$

For any $i \geq 1$, let $J_{i+1} = \mathbb{1}\{b^\pi_{i+1} \in B_{\gamma_1}\}$. Then, we have

$$\begin{aligned}
&\text{Regret}(\pi, I; \gamma_0) + \text{Regret}(\pi, I; \gamma_1) \\
&\geq q \sum_{i=1}^{I-1} \left( \mathbb{E}\left[ r(b^\pi_{i+1}, \gamma_0) - r(b^*(\gamma_0), \gamma_0) \right] + \mathbb{E}\left[ r(b^\pi_{i+1}, \gamma_1) - r(b^*(\gamma_1), \gamma_1) \right] \right) \\
&\geq q \frac{1}{3(24^2)\sqrt{I}} \sum_{i=1}^{I-1} \left( Q_i^{\pi,\gamma_0}\{b^\pi_{i+1} \notin B_{\gamma_0}\} + Q_i^{\pi,\gamma_1}\{b^\pi_{i+1} \notin B_{\gamma_1}\} \right) \\
&\geq q \frac{1}{3(24^2)\sqrt{I}} \sum_{i=1}^{I-1} \left( Q_i^{\pi,\gamma_0}\{b^\pi_{i+1} \in B_{\gamma_1}\} + Q_i^{\pi,\gamma_1}\{b^\pi_{i+1} \notin B_{\gamma_1}\} \right) \text{ [since } B_{\gamma_0} \text{ and } B_{\gamma_1} \text{ are disjoint]} \\
&= q \frac{1}{3(24^2)\sqrt{I}} \sum_{i=1}^{I-1} \left( Q_i^{\pi,\gamma_0}\{J_{i+1} = 1\} + Q_i^{\pi,\gamma_1}\{J_{i+1} = 0\} \right) \\
&\geq q \frac{1}{3(24^2)\sqrt{I}} \frac{1}{2} \sum_{i=1}^{I-1} \exp(-\mathcal{K}(Q_i^{\pi,\gamma_0}; Q_i^{\pi,\gamma_1})) \text{ [by Lemma A.7]} \\
&\geq q \frac{1}{3(24^2)\sqrt{I}} \frac{I-1}{2} \exp(-\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1})) \text{ [since } \mathcal{K}(Q_i^{\pi,\gamma_0}; Q_i^{\pi,\gamma_1}) \text{ is non-decreasing in } i] \\
&\geq q \frac{\sqrt{I}}{12(24^2)} \exp(-\mathcal{K}(Q_I^{\pi,\gamma_0}; Q_I^{\pi,\gamma_1})).
\end{aligned}$$

■

## Appendix F Proof of Lemma A.4

We first derive some preliminary results that will be useful in establishing Lemma A.4.

Consider the instance of problem $(P)$ defined in Theorem 3. Also, recall from Section 3.3 that $\gamma_0 = 1/2$. We make the following observations:

- $p(\gamma, b) \in [1/8, 7/8]$ for any $b \in B$ and $\gamma \in \Gamma^{(0)}$ and hence

$$p(\gamma, b)(1 - p(\gamma, b)) \geq 7/64. \tag{A-14}$$

- Let $b^\pi_{i,1}, b^\pi_{i,2}$ denote the bid prices in season $i$ under policy $\pi$. Let $r_2(b, \gamma) = (b - e)p(\gamma, b) = (b - 2)(1/2 - \gamma + \gamma b)$. Then, $b_2^*(\gamma) = 3/2 - 1/(4\gamma)$ is the optimal solution that minimizes $r_2(b, \gamma)$. Let $r_1(b, \gamma) = (b - e - \Delta V_{i,2}(1; \gamma))p(\gamma, b) = (b - 2 - \Delta V_{i,2}(1; \gamma))(1/2 - \gamma + \gamma b)$, where $\Delta V_{i,2}(1; \gamma) = q r_2(b_2^*(\gamma), \gamma) = -\frac{q}{\gamma}(\frac{\gamma}{2} + \frac{1}{4})^2$. Then $b_1^*(\gamma) = \frac{3\gamma - 1/2 - q\gamma^2/4 - q/16 - q\gamma/4}{2\gamma}$ is the optimal solution that minimizes $r_1(b, \gamma)$. We have

$$\text{Regret}(\pi, I; \gamma^{(0)})$$

$$
\begin{aligned}
&= \sum_{i=1}^{I} \mathbb{E}\big[q(b_{i,1}^{\pi} - e - \Delta V_{i,2}^{\pi}(1;\gamma))p(\gamma^{(0)}, b_{i,1}^{\pi}) + q(b_{i,2}^{\pi} - e)p(\gamma^{(0)}, b_{i,2}^{\pi})\big] - \\
&\quad \sum_{i=1}^{I} [q \min_{b\in B}\{(b - e - \Delta V_{i,2}(1;\gamma))p(\gamma^{(0)}, b)\} + q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\}] \\
&= \sum_{i=1}^{I} \mathbb{E}\big[q(b_{i,1}^{\pi} - e - q(b_{i,2}^{\pi} - e)p(\gamma^{(0)}, b_{i,2}^{\pi}))p(\gamma^{(0)}, b_{i,1}^{\pi}) + q(b_{i,2}^{\pi} - e)p(\gamma^{(0)}, b_{i,2}^{\pi})\big] - \\
&\quad I[q \min_{b\in B}\{(b - e - q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\})p(\gamma^{(0)}, b)\} + q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\}] \\
&= \sum_{i=1}^{I} \mathbb{E}\big[q(b_{i,1}^{\pi} - e - q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\})p(\gamma^{(0)}, b_{i,1}^{\pi}) + q(b_{i,2}^{\pi} - e)p(\gamma^{(0)}, b_{i,2}^{\pi})\big] - \\
&\quad I[q \min_{b\in B}\{(b - e - q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\})p(\gamma^{(0)}, b)\} + q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\}] + \\
&\quad \sum_{i=1}^{I} \mathbb{E}\big[q(q \min_{b\in B}\{(b-e)p(\gamma^{(0)}, b)\} - q(b_{i,2}^{\pi} - e)p(\gamma^{(0)}, b_{i,2}^{\pi}))p(\gamma^{(0)}, b_{i,1}^{\pi})\big] \\
&= q\sum_{i=1}^{I} \mathbb{E}\big[r_1(b_{i,1}^{\pi}, \gamma^{(0)}) - r_1(b_1^*(\gamma^{(0)}), \gamma^{(0)})\big] + \\
&\quad q\sum_{i=1}^{I} \mathbb{E}\big[(r_2(b_{i,2}^{\pi}, \gamma^{(0)}) - r_2(b_2^*(\gamma^{(0)}), \gamma^{(0)}))(1 - qp(\gamma^{(0)}, b_{i,1}^{\pi}))\big] \\
&\geq \frac{q}{8}\sum_{i=1}^{I}\sum_{t=1}^{2} \mathbb{E}\big[r_t(b_{i,t}^{\pi}, \gamma^{(0)}) - r_t(b_t^*(\gamma^{(0)}), \gamma^{(0)})\big]. \qquad \text{(A-15)}
\end{aligned}
$$

- $b_2^*(\gamma_0) = 1$ and $b_1^*(\gamma_0) = 1 - q/4$.
- Since $\gamma \leq 1$, we have

$$|b_2^*(\gamma) - b_2^*(\gamma_0)| = \frac{|\gamma - \gamma_0|}{4\gamma\gamma_0} \geq \frac{|\gamma - \gamma_0|}{2}.$$

- The absolute difference $|p(\gamma_0, b) - p(\gamma, b)|$ satisfies:

$$|p(\gamma_0, b) - p(\gamma, b)| = |\gamma - \gamma_0||b - 1| = |\gamma - \gamma_0||b - b_2^*(\gamma_0)| = |\gamma - \gamma_0||b - b_1^*(\gamma_0) - q/4|.$$

- $\frac{\partial^2 r_2(b,\gamma)}{\partial b^2} = 2\gamma \geq 2/3$. Using this, along with the fact that $\frac{\partial r_2(b,\gamma)}{\partial b} = 0$ at $b = b_2^*(\gamma)$, we obtain

$$
\begin{aligned}
r_2(b,\gamma) - r_2(b_2^*(\gamma),\gamma) &= \left| \int_{b_2^*(\gamma)}^{b} \int_{b_2^*(\gamma)}^{v} \frac{\partial^2 r_2(\tilde{b},\gamma)}{\partial \tilde{b}^2} d\tilde{b} dv \right| \\
&\geq \frac{2}{3} \left| \int_{b_2^*(\gamma)}^{b} \int_{b_2^*(\gamma)}^{v} d\tilde{b} dv \right| \\
&= \frac{1}{3}(b - b_2^*(\gamma))^2. \qquad \text{(A-16)}
\end{aligned}
$$

Similarly, we have

$$r_1(b,\gamma) - r_1(b_1^*(\gamma),\gamma) \geq \frac{1}{3}(b - b_1^*(\gamma))^2, \qquad \text{(A-17)}$$

and

$$r_1(b,\gamma) - r_1(b_1^*(\gamma),\gamma) \leq (b - b_1^*(\gamma))^2 \leq 1. \qquad \text{(A-18)}$$

**Lemma A.8** *For any $\gamma \in \Gamma^{(0)}$, $i \geq 1$, and policy $\pi$, we have*

$$\begin{aligned}
&\sum_{\boldsymbol{x_{i-1,2}} \in \mathbb{X}^{2i-2}} Q^{\pi,\gamma_0}_{i-1,2}(\boldsymbol{x_{i-1,2}}) \sum_{x_{i,1} \in \mathbb{X}} Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}}) \log\left(\frac{Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}{Q^{\pi,\gamma}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}\right) \\
\leq\ & \frac{248}{7} q(\gamma_0-\gamma)^2 \sqrt{\mathbb{E}\left[r_1(b^\pi_{i,1},\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)\right]} + \frac{4}{7}q^3(\gamma_0-\gamma)^2,
\end{aligned}$$

*and*

$$\begin{aligned}
&\sum_{\boldsymbol{x_{i,1}} \in \mathbb{X}^{2i-1}} Q^{\pi,\gamma_0}_{i,1}(\boldsymbol{x_{i,1}}) \sum_{x_{i,2} \in \mathbb{X}} Q^{\pi,\gamma_0}_{i,2}(x_{i,2}|\boldsymbol{x_{i,1}}) \log\left(\frac{Q^{\pi,\gamma_0}_{i,2}(x_{i,2}|\boldsymbol{x_{i,1}})}{Q^{\pi,\gamma}_{i,2}(x_{i,2}|\boldsymbol{x_{i,1}})}\right) \\
\leq & \frac{248}{7} q(\gamma_0-\gamma)^2 \sqrt{\mathbb{E}\left[r_2(b^\pi_{i,2},\gamma_0) - r_2(b^*_2(\gamma_0),\gamma_0)\right]} + \frac{4}{7}q^3(\gamma_0-\gamma)^2.
\end{aligned}$$

**Proof of Lemma A.8:** We first derive an upper bound on $\sum_{x_{i,1}\in\mathbb{X}} Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}}) \log\left(\frac{Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}{Q^{\pi,\gamma}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}\right)$:

$$\begin{aligned}
&\sum_{x_{i,1}\in\mathbb{X}} Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}}) \log\left(\frac{Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}{Q^{\pi,\gamma}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}\right) \\
=\ & qp(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})) \log\left(\frac{p(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))}{p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))}\right) + q(1-p(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))) \log\left(\frac{1-p(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))}{1-p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))}\right) \\
\leq\ & q\frac{(p(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})) - p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})))^2}{p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2}))(1-p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})))} \\
\leq\ & \frac{64}{7} q(p(\gamma_0, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})) - p(\gamma, b^\pi_{i,1}(\boldsymbol{x}_{i-1,2})))^2 \\
=\ & \frac{64}{7} q(\gamma_0-\gamma)^2 (b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}) - b^*_1(\gamma_0) - q/4)^2 \\
\leq\ & \frac{192}{7} q(\gamma_0-\gamma)^2 (r_1(b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}),\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)) + \frac{4}{7}q^3(\gamma_0-\gamma)^2 + \frac{32}{7}q^2(\gamma_0-\gamma)^2 |b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}) - b^*_1(\gamma_0)| \\
\leq\ & \frac{192}{7} q(\gamma_0-\gamma)^2 (r_1(b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}),\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)) + \frac{4}{7}q^3(\gamma_0-\gamma)^2 + \\
& \frac{32\sqrt{3}}{7} q^2(\gamma_0-\gamma)^2 \sqrt{r_1(b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}),\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)} \\
\leq & \frac{248}{7} q(\gamma_0-\gamma)^2 \sqrt{r_1(b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}),\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)} + \frac{4}{7}q^3(\gamma_0-\gamma)^2.
\end{aligned}$$

The first inequality holds by Lemma A.5. The second inequality holds by (A-14). The third and fourth inequalities hold by (A-17). The last inequality holds by (A-18).

Then, we have

$$\begin{aligned}
&\sum_{\boldsymbol{x_{i-1,2}} \in \mathbb{X}^{2i-2}} Q^{\pi,\gamma_0}_{i-1,2}(\boldsymbol{x_{i-1,2}}) \sum_{x_{i,1} \in \mathbb{X}} Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}}) \log\left(\frac{Q^{\pi,\gamma_0}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}{Q^{\pi,\gamma}_{i,1}(x_{i,1}|\boldsymbol{x_{i-1,2}})}\right) \\
\leq\ & \frac{248}{7} q(\gamma_0-\gamma)^2 \sum_{\boldsymbol{x_{i-1,2}} \in \mathbb{X}^{2i-2}} Q^{\pi,\gamma_0}_{i-1,2}(\boldsymbol{x_{i-1,2}}) \sqrt{r_1(b^\pi_{i,1}(\boldsymbol{x_{i-1,2}}),\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)} + \frac{4}{7}q^3(\gamma_0-\gamma)^2 \\
=\ & \frac{248}{7} q(\gamma_0-\gamma)^2 \mathbb{E}\left[\sqrt{r_1(b^\pi_{i,1},\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)}\right] + \frac{4}{7}q^3(\gamma_0-\gamma)^2 \\
\leq\ & \frac{248}{7} q(\gamma_0-\gamma)^2 \sqrt{\mathbb{E}\left[r_1(b^\pi_{i,1},\gamma_0) - r_1(b^*_1(\gamma_0),\gamma_0)\right]} + \frac{4}{7}q^3(\gamma_0-\gamma)^2.
\end{aligned}$$

By Lemma A.6, we have

$$\begin{aligned}
&\sum_{\boldsymbol{x_{i,1}}\in\mathbb{X}^{2i-1}} Q_{i,1}^{\pi,\gamma_0}(\boldsymbol{x_{i,1}}) \sum_{x_{i,2}\in\mathbb{X}} Q_{i,2}^{\pi,\gamma_0}(x_{i,2}|\boldsymbol{x_{i,1}}) \log\left(\frac{Q_{i,2}^{\pi,\gamma_0}(x_{i,2}|\boldsymbol{x_{i,1}})}{Q_{i,2}^{\pi,\gamma}(x_{i,2}|\boldsymbol{x_{i,1}})}\right)\\
\le&\frac{192}{7}q(\gamma_0-\gamma)^2\mathbb{E}\left[r_2(b_{i,2}^{\pi},\gamma_0)-r_2(b_2^*(\gamma_0),\gamma_0)\right]\\
\le&\frac{248}{7}q(\gamma_0-\gamma)^2\sqrt{\mathbb{E}\left[r_2(b_{i,2}^{\pi},\gamma_0)-r_2(b_2^*(\gamma_0),\gamma_0)\right]}+\frac{4}{7}q^3(\gamma_0-\gamma)^2.
\end{aligned}$$

■

**Proof of Lemma A.4:** (i) We apply the Chain rule for KL divergence (Theorem 2.5.3 in Cover and Thomas 2012):

$$\begin{aligned}
&\mathcal{K}(Q_{I,2}^{\pi,\gamma_0};Q_{I,2}^{\pi,\gamma})\\
=&\sum_{i=1}^{I}\sum_{\boldsymbol{x_{i-1,2}}\in\mathbb{X}^{2i-2}} Q_{i-1,2}^{\pi,\gamma_0}(\boldsymbol{x_{i-1,2}}) \sum_{x_{i,1}\in\mathbb{X}} Q_{i,1}^{\pi,\gamma_0}(x_{i,1}|\boldsymbol{x_{i-1,2}}) \log\left(\frac{Q_{i,1}^{\pi,\gamma_0}(x_{i,1}|\boldsymbol{x_{i-1,2}})}{Q_{i,1}^{\pi,\gamma}(x_{i,1}|\boldsymbol{x_{i-1,2}})}\right)+\\
&\sum_{i=1}^{I}\sum_{\boldsymbol{x_{i,1}}\in\mathbb{X}^{2i-1}} Q_{i,1}^{\pi,\gamma_0}(\boldsymbol{x_{i,1}}) \sum_{x_{i,2}\in\mathbb{X}} Q_{i,2}^{\pi,\gamma_0}(x_{i,2}|\boldsymbol{x_{i,1}}) \log\left(\frac{Q_{i,2}^{\pi,\gamma_0}(x_{i,2}|\boldsymbol{x_{i,1}})}{Q_{i,2}^{\pi,\gamma}(x_{i,2}|\boldsymbol{x_{i,1}})}\right)\\
\le&\frac{248}{7}q(\gamma_0-\gamma)^2\sum_{i=1}^{I}\left(\sqrt{\mathbb{E}\left[r_1(b_{i,1}^{\pi},\gamma_0)-r_1(b_1^*(\gamma_0),\gamma_0)\right]}+\sqrt{\mathbb{E}\left[r_2(b_{i,2}^{\pi},\gamma_0)-r_2(b_2^*(\gamma_0),\gamma_0)\right]}\right)\\
&\frac{8}{7}q^3I(\gamma_0-\gamma)^2\\
\le&\frac{248}{7}q(\gamma_0-\gamma)^2\sqrt{2I}\sqrt{\sum_{i=1}^{I}\left(\mathbb{E}\left[r_1(b_{i,1}^{\pi},\gamma_0)-r_1(b_1^*(\gamma_0),\gamma_0)\right]+\mathbb{E}\left[r_2(b_{i,2}^{\pi},\gamma_0)-r_2(b_2^*(\gamma_0),\gamma_0)\right]\right)}+\\
&\frac{8}{7}q^3I(\gamma_0-\gamma)^2\\
\le&\frac{992}{7}\sqrt{q}(\gamma_0-\gamma)^2\sqrt{I}\sqrt{\mathrm{Regret}(\pi,I;\gamma_0)}+\frac{8}{7}q^3I(\gamma_0-\gamma)^2
\end{aligned}$$

The first inequality holds by Lemma A.8. The last inequality holds by (A-15).

For $\gamma_1=\gamma_0+\frac{1}{4}I^{-2/7}$ and $q=K_4I^{-1/7}$, we have

$$\sqrt{\mathrm{Regret}(\pi,I;\gamma_0)}\ge\frac{7}{992\sqrt{qI}(\gamma_0-\gamma_1)^2}\mathcal{K}(Q_{I,2}^{\pi,\gamma_0};Q_{I,2}^{\pi,\gamma_1})-\frac{8q^{5/2}\sqrt{I}}{992}=\frac{7I^{1/7}}{62\sqrt{K_4}}\mathcal{K}(Q_{I,2}^{\pi,\gamma_0};Q_{I,2}^{\pi,\gamma_1})-\frac{(K_4)^{5/2}I^{1/7}}{124}.$$

(ii) We first define two intervals $B_{\gamma_0}\subset B$ and $B_{\gamma_1}\subset B$:

$$B_{\gamma_0}:=\left\{b:|b_2^*(\gamma_0)-b|\le\frac{1}{24I^{2/7}}\right\}\text{ and }B_{\gamma_1}:=\left\{b:|b_2^*(\gamma_1)-b|\le\frac{1}{24I^{2/7}}\right\}.$$

Note that $B_{\gamma_0}$ and $B_{\gamma_1}$ are disjoint since $|b_2^*(\gamma_1)-b_2^*(\gamma_0)|\ge\frac{|\gamma_1-\gamma_0|}{2}=\frac{1}{8I^{2/7}}$. Recall from (A-16) that $r_2(b_2^*(\gamma),\gamma)-r_2(b,\gamma)\ge\frac{1}{3}(b_2^*(\gamma)-b)^2$. For each $\gamma\in\{\gamma_0,\gamma_1\}$, if $b\notin B_\gamma$, then

$$r_2(b_2^*(\gamma),\gamma)-r_2(b,\gamma)\ge\frac{1}{3}(b_2^*(\gamma)-b)^2\ge\frac{1}{3(24^2)I^{4/7}}.$$

For any $i\ge1$, let $J_{i+1}=\mathbb{1}\{b_{i+1,2}^{\pi}\in B_{\gamma_1}\}$. Then, we have

$$\mathrm{Regret}(\pi,I;\gamma_0)+\mathrm{Regret}(\pi,I;\gamma_1)$$

$$
\begin{aligned}
&\geq \frac{q}{8}\sum_{i=1}^{I-1}\left(\mathbb{E}\left[r_2(b^{\pi}_{i+1,2},\gamma_0)-r_2(b^{*}_2(\gamma_0),\gamma_0)\right]+\mathbb{E}\left[r_2(b^{\pi}_{i+1,2},\gamma_1)-r_2(b^{*}_2(\gamma_1),\gamma_1)\right]\right)\\
&\geq q\frac{1}{24^3 I^{4/7}}\sum_{i=1}^{I-1}\left(Q^{\pi,\gamma_0}_{i,2}\{b^{\pi}_{i+1,2}\notin B_{\gamma_0}\}+Q^{\pi,\gamma_1}_{i,2}\{b^{\pi}_{i+1}\notin B_{\gamma_1}\}\right)\\
&\geq q\frac{1}{24^3 I^{4/7}}\sum_{i=1}^{I-1}\left(Q^{\pi,\gamma_0}_{i,2}\{b^{\pi}_{i+1,2}\in B_{\gamma_1}\}+Q^{\pi,\gamma_1}_{i,2}\{b^{\pi}_{i+1,2}\notin B_{\gamma_1}\}\right)\ \text{[since } B_{\gamma_0}\text{ and } B_{\gamma_1}\text{ are disjoint]}\\
&= q\frac{1}{24^3 I^{4/7}}\sum_{i=1}^{I-1}\left(Q^{\pi,\gamma_0}_{i,2}\{J_{i+1}=1\}+Q^{\pi,\gamma_1}_{i,2}\{J_{i+1}=0\}\right)\\
&\geq q\frac{1}{24^3 I^{4/7}}\frac{1}{2}\sum_{i=1}^{I-1}\exp(-\mathcal{K}(Q^{\pi,\gamma_0}_{i,2};Q^{\pi,\gamma_1}_{i,2}))\ \text{[by Lemma A.7]}\\
&\geq q\frac{1}{24^3 I^{4/7}}\frac{I-1}{2}\exp(-\mathcal{K}(Q^{\pi,\gamma_0}_{I,2};Q^{\pi,\gamma_1}_{I,2}))\ \text{[since } \mathcal{K}(Q^{\pi,\gamma_0}_{i,2};Q^{\pi,\gamma_1}_{i,2})\text{ is non-decreasing in } i]\\
&\geq \frac{K_4 I^{2/7}}{4(24^3)}\exp(-\mathcal{K}(Q^{\pi,\gamma_0}_{I,2};Q^{\pi,\gamma_1}_{I,2})).
\end{aligned}
$$

■

## Appendix G Further Results on the Regret

In this section, we first establish an upper bound on the regret under *any* policy that is linear in $T$ (i.e., the number of periods in a season) and independent of $L$ (i.e., the number of geographical locations from where the impressions are acquired). Then, we consider the following setting of the mobile-promotion platform's problem: All impressions arrive from a single location that has the win curve $p(\gamma,b)=\exp(\gamma(b-e))$, and the start times and the end times of the campaigns in each season are ordered in the same way; that is, the campaigns end in the order of their arrival. For this setting, we show that in Theorem A.1 that the regret under our policy is $\mathcal{O}(\sqrt{T}\log^2(T))$. This result also allows us to analyze the setting where campaigns can start and finish in different seasons (Theorem A.2).

Let $\mathcal{P}=\{p_1(\gamma,b),\cdots,p_\upsilon(\gamma,b)\}$ denote the set of possible functional forms of the win curves, where $\upsilon$ is independent of $L$. Thus, the win curve at each location has its own unknown parameters but the set of possible functional forms of the win curves is fixed and limited. Let $\Gamma_j$ denote the set of all possible values of $\gamma$ for $p_j(\gamma,b)$. Then, we have

**Lemma A.9** *For any policy $\pi$, there exists an upper bound on the regret that is linear in $T$ and independent of $L$. Precisely, Regret$(\pi,I;\gamma^{(0)})\leq K_0 IT(b^{\max}-b^{\min})^2$, and the constant $K_0$ is independent of $I$, $T$, and $L$.*

**Proof of Lemma A.9**: Using Lemma 2, we have

$$
\begin{aligned}
\text{Regret}(\pi,I;\gamma^{(0)}) &\leq K_0\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}q_l\left(b^{\pi}_{i,t,l}(X^{\pi}_{i,1},\hat{X}^{\pi}_{i,t})-b^{*}_{i,t,l}(\boldsymbol{c_i}(\hat{X}^{\pi}_{i,t});\gamma^{(0)})\right)^2\right]\\
&\leq K_0 IT(b^{\max}-b^{\min})^2.
\end{aligned}
$$

Recall from the proof of Lemma A.1 that $K_0$ is independent of $T$ and

$$K_0 \leq \sup_{j\in\{1,\cdots,v\}} \sup_{(\alpha,\gamma,b)\in[b^{\min}-e,0]\times\Gamma_j\times B} \frac{\partial^2 f_j(b,\alpha,\gamma)}{\partial b^2}/2,$$

where $f_j(b,\alpha,\gamma) = p_j(\gamma,b)(b-e-\alpha)$ is continuous in $b$, $\alpha$ and $\gamma$. It is clear that $K_0$ is independent of $L$ and $T$. Consequently, the upper bound on the regret is independent of $L$ and

$$\text{Regret}(\pi, I; \gamma^{(0)}) = \mathcal{O}(T). \qquad \blacksquare$$

Next, we establish an $\mathcal{O}(\sqrt{T}\log^2(T))$ upper bound on the regret under our policy for a special setting.

**Theorem A.1** *Consider the following setting of the mobile-promotion platform's problem: All impressions arrive from a single location and the win curve at that location is $p(\gamma,b) = \exp(\gamma(b-e))$ for $b \in B$ and $\gamma \in \Gamma^{(0)} = [\gamma^{\min}, \gamma^{\max}]$, where $\gamma^{\max} \geq \gamma^{\min} > 0$. The start times and the end times of the campaigns in each season are ordered in the same way; that is, the campaigns end in the order of their arrival. Then, for $T \geq 3$, there exists a constant $K_{10} > 0$ that is independent of $I$ and $T$, such that*

$$\textit{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq K_{10}\sqrt{IT}\log^2(T).$$

**Proof of Theorem A.1:** Recall from equation (9) in the proof of Theorem 1 that

$$\text{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq \left(\sqrt{\frac{2IT}{L}} + 1\right) K_0(\hat{K} + \check{K}),$$

where $K_0$ and $\hat{K}$ are independent of $I$ and $T$. We show that $\left|b^*_{i,t}(c;\gamma) - b^*_{i,t}(c;\hat{\gamma})\right| \leq \frac{2\log T}{(\gamma^{\min})^2}|\gamma-\hat{\gamma}|$ in Lemma A.10 below. Thus, $\check{K} = 4\log^2(T)LK_{mle}/(\gamma^{\min})^4$ and $\text{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq K_{10}\sqrt{IT}\log^2(T)$ for $K_{10} = \left(\sqrt{\frac{2}{L}}+1\right)K_0\left(\hat{K} + \frac{4LK_{mle}}{(\gamma^{\min})^4}\right)$. $\blacksquare$

We now prepare the groundwork to state and prove Lemma A.10. Consider the setting defined in Theorem A.1. Recall that $c_{i,t,j}$ is the number of unmet impressions for campaign $(i,j)$ at the beginning of period $(i,t)$. Then, the optimization problem of the clairvoyant problem for season $i$ can be written as the following DP, in which the state in any period $(i,t)$ is the total number of unmet impressions $c_{i,t}$ over all the campaigns that end in or after that period; i.e., $c_{i,t} = \sum_{j:\underline{t}_{i,j}\geq t} c_{i,t,j}$. A formal description of this DP follows. Let $C_{i,t} := \sum_{j:\underline{t}_{i,j}\geq t} W_{i,j}$ denote the total number of required impressions for all the campaigns in season $i$ that end in or after period $(i,t)$, thus, $c_{i,t} \leq C_{i,t}$. Let $\hat{C}_{i,t} := \sum_{j:\bar{t}_{i,j}\geq t} W_{i,j}$ denote the total number of required impressions over all the campaigns that start in or after period $(i,t)$, thus, $c_{i,t} \geq \hat{C}_{i,t+1}$. If $c_{i,t} = \hat{C}_{i,t+1}$, then there is no active campaign in period $(i,t)$, and thus no bid should be placed in that period.

Next, we present some observations when $c_{i,t} > \hat{C}_{i,t+1}$; these will be used in defining the DP recursion for season $i$.

- If $c_{i,t} > C_{i,t+1}$, then the optimal allocation policy assigns the impression won (if applicable) in period $(i,t)$ to the lowest-indexed active campaign that ends in that period. As a result, no impression is assigned to

campaigns that end after period $(i,t)$; thus, the total number of unmet impressions for these campaigns in period $t+1$ is their total number of required impressions, i.e., $C_{i,t+1}$. Thus, if $c_{i,t} > C_{i,t+1}$, then the state in period $(i,t+1)$ is $c_{i,t+1} = C_{i,t+1}$, regardless of whether or not the platform wins an impression in period $(i,t)$.

- If $c_{i,t} \le C_{i,t+1}$, then the state in period $(i,t+1)$ is either $c_{i,t+1} = c_{i,t} - 1$ or $c_{i,t+1} = c_{i,t}$.
  - If, in period $(i,t)$, an impression arrives from one of the locations in $\mathcal{L}$, say $l$, and it is won, then the state in period $(i,t+1)$ is $c_{i,t+1} = c_{i,t} - 1$.
  - If, in period $(i,t)$, no impression arrives or the arriving impression is not won, then the state in period $(i,t+1)$ is $c_{i,t+1} = c_{i,t}$.

We now formulate the DP. In the sequel, we drop the time index $(i,t)$ of $c_{i,t}$ when there is no ambiguity. Let $V_{i,t}(c;\gamma)$ denote the optimal cost-to-go function of the DP and let $b^*_{i,t}(c;\gamma)$ denote the optimal bidding amount in period $(i,t)$ when there are $c$ unmet impressions over the campaigns that end in or after period $(i,t)$. If $c = \hat{C}_{i,t+1}$, then there are no active campaigns, and thus no bid should be placed (i.e., $b^*_{i,t}(c;\gamma) = 0$) and $V_{i,t}(c;\gamma) = V_{i,t+1}(c;\gamma)$. Otherwise, for $c > \hat{C}_{i,t+1}$, $V_{i,t}(c;\gamma)$ satisfies the following recursion:

For any $c > \hat{C}_{i,t+1}$,

$$\begin{aligned} V_{i,t}(c;\gamma) = \quad & \min_{b\in B} \left\{ \begin{array}{l} \mathbb{1}\{c > C_{i,t+1}\}\left[qp(\gamma,b)(b-e) + V_{i,t+1}(C_{i,t+1};\gamma)\right] + \\ \mathbb{1}\{c \le C_{i,t+1}\} qp(\gamma,b)[b - e + V_{i,t+1}(c-1;\gamma)] + \\ \mathbb{1}\{c \le C_{i,t+1}\}\left[1 - qp(\gamma,b)\right] V_{i,t+1}(c;\gamma) \end{array} \right\} \\ = \quad & q \min_{b\in B} p(\gamma,b)\left[b - e - \mathbb{1}\{c \le C_{i,t+1}\}\Delta V_{i,t+1}(c;\gamma)\right] + \\ & \mathbb{1}\{c > C_{i,t+1}\} V_{i,t+1}(C_{i,t+1};\gamma) + \mathbb{1}\{c \le C_{i,t+1}\} V_{i,t+1}(c;\gamma), \end{aligned}$$

where $C_{i,T+1} = 0$, $V_{i,T+1}(c;\gamma) = 0$ for any $c \ge 0$, and

$$\Delta V_{i,t+1}(c;\gamma) := V_{i,t+1}(c;\gamma) - V_{i,t+1}(c-1;\gamma).$$

The optimal bidding amount when an impression arrives is as follows:

For any $c > \hat{C}_{i,t+1}$,

$$b^*_{i,t}(c;\gamma) = \arg\min_{b\in B} p(\gamma,b)\left[b - e - \mathbb{1}\{c \le C_{i,t+1}\}\Delta V_{i,t+1}(c;\gamma)\right]. \tag{A-19}$$

**Lemma A.10** *Consider the setting defined in Theorem A.1. For all $\gamma, \hat{\gamma} \in \Gamma^{(0)} = [\gamma^{\min}, \gamma^{\max}]$, $c > \hat{C}_{i,t+1}$, $1 \le i \le I$, and $1 \le t \le T$, we have*

$$\left|b^*_{i,t}(c;\gamma) - b^*_{i,t}(c;\hat{\gamma})\right| \le \frac{2\log T}{(\gamma^{\min})^2}|\gamma - \hat{\gamma}|.$$

To establish Lemma A.10, we compare the optimal bidding amount defined in (A-19) with the one under the setting where there is only one campaign in each season that starts at the beginning of the season and ends at the end the season. When there is only one campaign in each season, the optimization problem of the clairvoyant problem for season $i$ can be written as the following DP, in which the state in any period $(i,t)$ is the total number of unmet impressions $c_{i,t}$ of the campaign. In the sequel, we drop the time index $(i,t)$ of $c_{i,t}$ when there is no ambiguity. Let $\tilde{V}_{i,t}(c;\gamma)$ denote the optimal cost-to-go function of the DP and let $\tilde{b}_{i,t}(c;\gamma)$ denote the optimal bidding amount in period $(i,t)$ when there are $c$ unmet impressions. If $c=0$, then no bid should be placed (i.e., $\tilde{b}_{i,t}(0;\gamma)=0$) and $\tilde{V}_{i,t}(0;\gamma)=0$. Otherwise, for $c>0$, $\tilde{V}_{i,t}(c;\gamma)$ satisfies the following recursion:

$$\tilde{V}_{i,t}(c;\gamma) = q \min_{b\in B} p(\gamma,b)\left[b - e - \Delta\tilde{V}_{i,t+1}(c;\gamma)\right] + \tilde{V}_{i,t+1}(c;\gamma). \tag{A-20}$$

where $\Delta\tilde{V}_{i,t+1}(c;\gamma) := \tilde{V}_{i,t+1}(c;\gamma) - \tilde{V}_{i,t+1}(c-1;\gamma)$. The optimal bidding amount when an impression arrives is $\tilde{b}_{i,t}(c;\gamma) = \arg\min_{b\in B} p(\gamma,b)\left[b - e - \Delta\tilde{V}_{i,t+1}(c;\gamma)\right]$.

**Lemma A.11** *For $\hat{C}_{i,t} < c \le C_{i,t}$, we have $\Delta\tilde{V}_{i,t}(1;\gamma) \le \Delta V_{i,t}(c;\gamma)$.*

**Proof of Lemma A.11:** In the notation of Lemma A.1, let $\alpha = \Delta V_{i,t+1}(c;\gamma)$. The proof is by induction on $t$. For $t = T+1$, $\Delta\tilde{V}_{i,T+1}(1;\gamma) = \Delta V_{i,T+1}(c;\gamma) = 0$ for all $c>0$. For $2 \le t \le T$, suppose that $\Delta\tilde{V}_{i,t+1}(1;\gamma) \le \Delta V_{i,t+1}(c;\gamma)$ holds for all $\hat{C}_{i,t+1} < c \le C_{i,t+1}$. We show that $\Delta\tilde{V}_{i,t}(1;\gamma) \le \Delta V_{i,t}(c;\gamma)$ for all $\hat{C}_{i,t} < c \le C_{i,t}$ using the following three cases.

- Case 1: $c \le C_{i,t+1}$. In this case, we have

$$\begin{aligned}
&\Delta\tilde{V}_{i,t}(1;\gamma) - \Delta V_{i,t}(c;\gamma) \\
=\ &\Delta\tilde{V}_{i,t+1}(1;\gamma) + q\min_{b\in B} p(\gamma,b)[b - e - \Delta\tilde{V}_{i,t+1}(1;\gamma)] - \\
&\Delta V_{i,t+1}(c;\gamma) - q\min_{b\in B} p(\gamma,b)[b - e - \Delta V_{i,t+1}(c;\gamma)] + \\
&\mathbb{1}\{c > \hat{C}_{i,t}+1\} q\min_{b\in B} p(\gamma,b)[b - e - \Delta V_{i,t+1}(c-1;\gamma)].
\end{aligned} \tag{A-21}$$

  If $c > \hat{C}_{i,t+1}+1$, then we have

$$\begin{aligned}
&\Delta\tilde{V}_{i,t}(1;\gamma) - \Delta V_{i,t}(c;\gamma) \\
=\ &\Delta\tilde{V}_{i,t+1}(1;\gamma) + q\min_{b\in B} p(\gamma,b)[b - e - \Delta\tilde{V}_{i,t+1}(1;\gamma)] - \\
&\Delta V_{i,t+1}(c;\gamma) - q\min_{b\in B} p(\gamma,b)[b - e - \Delta V_{i,t+1}(c;\gamma)] + \\
&q\min_{b\in B} p(\gamma,b)[b - e - \Delta V_{i,t+1}(c-1;\gamma)] \\
\le\ &\Delta\tilde{V}_{i,t+1}(1;\gamma) + qp(\gamma,b^*(\alpha,\gamma))[b^*(\alpha,\gamma) - e - \Delta\tilde{V}_{i,t+1}(1;\gamma)] - \\
&\Delta V_{i,t+1}(c;\gamma) - qp(\gamma,b^*(\alpha,\gamma))[b^*(\alpha,\gamma) - e - \Delta V_{i,t+1}(c;\gamma)] \\
=\ &\left(1 - qp(\gamma,b^*(\alpha,\gamma))\right)\left(\Delta\tilde{V}_{i,t+1}(1;\gamma) - \Delta V_{i,t+1}(c;\gamma)\right)
\end{aligned}$$

$$\leq 0.$$

The first inequality holds, since

$$q \min_{b \in B} p(\gamma, b)[b - e - \Delta V_{i,t+1}(c-1;\gamma)] \leq qp(\gamma, b^{\min})[b^{\min} - e - \Delta V_{i,t+1}(c-1;\gamma)] \leq 0 \text{ [by Lemma A.2].}$$

The second inequality holds by the induction hypothesis.

If $c = \hat{C}_{i,t+1} + 1$, then we have

$$\begin{aligned}
& \Delta \tilde{V}_{i,t}(1;\gamma) - \Delta V_{i,t}(c;\gamma) \\
= & \Delta \tilde{V}_{i,t+1}(1;\gamma) + q \min_{b \in B} p(\gamma, b)[b - e - \Delta \tilde{V}_{i,t+1}(1;\gamma)] - \\
& \Delta V_{i,t+1}(c;\gamma) - q \min_{b \in B} p(\gamma, b)[b - e - \Delta V_{i,t+1}(c;\gamma)] \\
\leq & \ (1 - qp(\gamma, b^*(\alpha, \gamma))) \left(\Delta \tilde{V}_{i,t+1}(1;\gamma) - \Delta V_{i,t+1}(c;\gamma)\right) \\
\leq & \ 0.
\end{aligned}$$

The second inequality holds by the induction hypothesis.

- Case 2: $c = C_{i,t+1} + 1$. In this case, we have

$$\begin{aligned}
& \Delta \tilde{V}_{i,t}(1;\gamma) - \Delta V_{i,t}(c;\gamma) \\
= & \Delta \tilde{V}_{i,t}(1;\gamma) + q \min_{b \in B} p(\gamma, b)[b - e - \Delta \tilde{V}_{i,t+1}(1;\gamma)] - \\
& q \min_{b \in B} p(\gamma, b)(b - e) - V_{i,t+1}(C_{i,t+1};\gamma) + \\
& \mathbb{1}\{c > \hat{C}_{i,t} + 1\} q \min_{b \in B} p(\gamma, b)[b - e - \Delta V_{i,t+1}(c-1;\gamma)] + V_{i,t+1}(c-1;\gamma) \\
= & \Delta \tilde{V}_{i,t}(1;\gamma) + q \min_{b \in B} p(\gamma, b)[b - e - \Delta \tilde{V}_{i,t+1}(1;\gamma)] - \\
& q \min_{b \in B} p(\gamma, b)(b - e) + \mathbb{1}\{c > \hat{C}_{i,t} + 1\} q \min_{b \in B} p(\gamma, b)[b - e - \Delta V_{i,t+1}(c-1;\gamma)].
\end{aligned}$$

Following an argument similar to that in Case 1 by letting $\Delta V_{i,t+1}(c;\gamma) = 0$ in (A-21), we have $\Delta \tilde{V}_{i,t}(1;\gamma) \leq \Delta V_{i,t}(c;\gamma)$.

- Case 3: $c > C_{i,t+1} + 1$. In this case, we have

$$\Delta \tilde{V}_{i,t}(1;\gamma) - \Delta V_{i,t}(c;\gamma) = \Delta \tilde{V}_{i,t}(1;\gamma) \leq 0.$$

The inequality holds by Lemma A.2. ■

**Proof of Lemma A.10:** Consider the setting defined in Theorem A.1. Let $z_{i,t} = \gamma(b_{i,t} - e)$ and $p(z) = \exp(z)$ for $z \in Z = [\gamma(b^{\min} - e), \gamma(b^{\max} - e)]$. Let $b^{\pi}_{i,t}$ denote the bidding price in period $(i,t)$ under policy $\pi$. Then, the optimization problem defined in Theorem A.1, i.e.,

$$\min_{\pi \in \Pi} \ \mathbb{E}\left[\sum_{i=1}^{I} \sum_{t=1}^{T} (b^{\pi}_{i,t} - e) q d_{i,t}\right]$$

can be equivalently written as

$$\frac{1}{\gamma}\min_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{i,t}^{\pi} q d_{i,t}\right] \text{ or } \min_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{i,t}^{\pi} q d_{i,t}\right].$$

Note that, for the clairvoyant problem, the optimal decision in any period $(i,t)$ when there are $c$ unmet impressions over all the campaigns that end in or after that period, denoted by $z_{i,t}^*(c)$, is independent of the parameter $\gamma$. Then, we have $b_{i,t}^*(c;\gamma) = z_{i,t}^*(c)/\gamma + e$ or $z_{i,t}^*(c) = \gamma(b_{i,t}^*(c;\gamma) - e)$. Consequently, $\left|b_{i,t}^*(c;\gamma) - b_{i,t}^*(c;\hat{\gamma})\right| = \frac{|z_{i,t}^*(c)|}{\gamma\hat{\gamma}}|\gamma - \hat{\gamma}|$. Similarly, when there is only one campaign in each season, let $\tilde{z}_{i,t}(c) = \gamma(\tilde{b}_{i,t}(c;\gamma) - e)$ denote the optimal decision in period $(i,t)$ when there are $c$ unmet impressions. In the notion of Lemma A.1, $b_{i,t}^*(c;\gamma) = b^*(\Delta V_{i,t+1}(c),\gamma)$ and $\tilde{b}_{i,t}(c;\gamma) = b^*(\Delta\tilde{V}_{i,t+1}(c),\gamma)$. Recall that $\Delta\tilde{V}_{i,t}(1;\gamma) \le \Delta V_{i,t}(c;\gamma)$ by Lemma A.11. In addition, it is straightforward that $\Delta\tilde{V}_{i,2}(1;\gamma) \le \Delta\tilde{V}_{i,t}(1;\gamma)$ for $t \ge 2$. Thus, we have $\Delta\tilde{V}_{i,2}(1;\gamma) \le \Delta\tilde{V}_{i,t}(1;\gamma) \le \Delta V_{i,t}(c;\gamma)$. Recall that $b^*(\alpha,\gamma)$ is increasing in $\alpha$ by part (iv) of Lemma A.1. Thus, $\tilde{b}_{i,1}(1;\gamma) \le b_{i,t}^*(c;\gamma)$ for all $t \ge 1$ and $c > \hat{C}_{i,t+1}$, which implies that $\tilde{z}_{i,1}(1) \le z_{i,t}^*(c)$. In addition, it is straightforward that $z_{i,t}^*(c) \le 0$. Thus, we have

$$\left|b_{i,t}^*(c;\gamma) - b_{i,t}^*(c;\hat{\gamma})\right| = \frac{|z_{i,t}^*(c)|}{\gamma\hat{\gamma}}|\gamma - \hat{\gamma}| \le \frac{|\tilde{z}_{i,1}(1)|}{(\gamma^{\min})^2}|\gamma - \hat{\gamma}|.$$

We show by backward induction below (under the title "Derivation of Inequality (A-22)") that

$$\tilde{z}_{i,t}(1) \ge -\log(T-t+1) - 1. \tag{A-22}$$

Thus, $|\tilde{z}_{i,1}(1)| \le 2\log T$ and $\left|b_{i,t}^*(c;\gamma) - b_{i,t}^*(c;\hat{\gamma})\right| \le \frac{2\log T}{(\gamma^{\min})^2}|\gamma - \hat{\gamma}|$. ■

**Derivation of Inequality** (A-22)**:** When there is only one campaign in each season, for any $c \ge 1$, the optimal cost-to-go function $\tilde{V}_{i,t}(c)$ in (A-20) satisfies the following recursion:

$$\gamma\tilde{V}_{i,t}(c) = \min_{z\in Z} qp(z)\left[z - \gamma\Delta\tilde{V}_{i,t+1}(c)\right] + \gamma\tilde{V}_{i,t+1}(c).$$

Then, the optimal decision $\tilde{z}_{i,t}(c)$ in period $(i,t)$ at state $c$ is

$$\tilde{z}_{i,t}(c) = \arg\min_{z\in Z} p(z)\left[z - \gamma\Delta\tilde{V}_{i,t+1}(c)\right]. \tag{A-23}$$

Solving (A-23), we have

$$\tilde{z}_{i,t}(1) = \gamma\tilde{V}_{i,t+1}(1) - 1.$$

Combining with $\gamma\tilde{V}_{i,t}(1) = q\exp(\tilde{z}_{i,t}(1))[\tilde{z}_{i,t}(1) - \gamma\tilde{V}_{i,t+1}(1)] + \gamma\tilde{V}_{i,t+1}(1)$, we have $\tilde{z}_{i,T}(1) = -1$ and $\tilde{z}_{i,t-1}(1) = \tilde{z}_{i,t}(1) - q\exp(\tilde{z}_{i,t}(1))$. We show that $\tilde{z}_{i,t}(1) \ge -\log(T-t+1) - 1$ by backward induction. When $t = T$, we have

$$\tilde{z}_{i,T}(1) = -1 = -\log(T-T+1) - 1.$$

Suppose $\tilde{z}_{i,t+1}(1) \ge -\log(T-t) - 1$. We show that $\tilde{z}_{i,t}(1) \ge -\log(T-t+1) - 1$:

$$\tilde{z}_{i,t}(1) = \tilde{z}_{i,t+1}(1) - q\exp(\tilde{z}_{i,t+1}(1))$$

$$
\begin{aligned}
&\geq \tilde{z}_{i,t+1}(1) - \exp(\tilde{z}_{i,t+1}(1)) \\
&\geq -\log(T-t) - 1 - \exp(-\log(T-t) - 1) \\
&\geq -\log(T-t+1) - 1.
\end{aligned}
$$

The second inequality holds since $z - \exp(z)$ increases in $z$ for all $z \leq 0$. Let $y = T - t \geq 1$. Showing the validity of the third inequality is equivalent to showing

$$
-\log y - 1 - \exp(-\log y - 1) \geq -\log(y+1) - 1 \Leftrightarrow \log(\frac{y+1}{y})y \geq \exp(-1).
$$

Since $\log(\frac{y+1}{y})y$ increases in $y$ for $y \geq 1$, we have

$$
\log(\frac{y+1}{y})y \geq \log 2 \geq \exp(-1). \qquad \blacksquare
$$

We now discuss the advantage offered by the single-location and exponential win-curve assumptions. To establish the upper bound $\mathcal{O}(\sqrt{T}\log^2 T)$ on the regret in Theorem A.1, we use a DP, which defines the optimal bidding amount for the clairvoyant problem, to derive an upper bound on the difference between the optimal bids under two arbitrary parameters of the win-curve. We show that this upper bound is $\mathcal{O}(\log T)$ times the absolute difference between the two parameters (Lemma A.10). The single-location assumption and the exponential win-curve assumption help us establish certain properties of the optimal bidding amount, which in turn help us prove the upper bound in Lemma A.10. More specifically, under the single-location and exponential win-curve assumptions, we can isolate the win-curve parameter from the optimization problem through a linear transformation of the optimal bidding amount to show the following property: For two arbitrary values $\gamma$ and $\hat{\gamma}$ of the win-curve parameter, the difference between the corresponding optimal bids, i.e., $|b^*_{i,t}(c;\gamma) - b^*_{i,t}(c;\hat{\gamma})|$, equals $\frac{|z^*_{i,t}(c)|}{\gamma\hat{\gamma}}|\gamma - \hat{\gamma}|$, where $z^*_{i,t}(c)$ is the optimal decision in a new problem that is independent of the win-curve parameter. This property, together with the exponential win-curve assumption and the monotonicity of the optimal bidding amount with respect to time period $t$ and the remaining number of unmet impressions $c$, helps us derive an $\mathcal{O}(\log T)$ upper bound on $|z^*_{i,t}(c)|$. In general (i.e., without the single-location and exponential win-curve assumptions), we are unable to establish this result.

Theorem A.1 also lets us address the setting where campaigns can start and finish in different seasons.

**Theorem A.2** *Consider the following setting of the mobile-promotion platform's problem: All impressions arrive from a single location and the win curve at that location is* $p(\gamma, b) = \exp(\gamma(b-e))$ *for* $b \in B$ *and* $\gamma \in \Gamma^{(0)} = [\gamma^{\min}, \gamma^{\max}]$*, where* $\gamma^{\max} \geq \gamma^{\min} > 0$*. The start times and the end times of the campaigns are ordered in the same way; that is, the campaigns end in the order of their arrival (the campaigns can start and end in different seasons). Then, for* $I \geq 3$ *and* $T \geq 3$*, we have*

$$
\mathit{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq K_{11}\sqrt{I}\log^2(I),
$$

*where* $K_{11} = 4K_{10}\sqrt{T}\log^2 T$ *for constant* $K_{10}$ *that is independent of* $I$ *and* $T$*.*

**Proof of Theorem A.2:** Consider the $I$ seasons as one dummy season consisting of $IT$ periods. Then, by Theorem A.1, we have

$$\text{Regret}(\textsc{BidAlloc}, I; \gamma^{(0)}) \leq K_{10}\sqrt{IT}\log^2(IT) \leq K_{11}\sqrt{I}\log^2(I),$$

where $K_{11} = 4K_{10}\sqrt{T}\log^2 T$. ■

## Appendix H The Well-Separated Case

In this section, we consider a "well-separated" setting defined in Broder and Rusmevichientong (2012); see Assumptions 5 and 6 below. We assume that all impressions arrive from a single location. In section H.1, we show that the regret under any policy is $\Omega(\log I)$ under the well-separated condition. Then, in section H.2, we propose a policy similar to the one presented in Broder and Rusmevichientong (2012), which achieves a matching upper bound on the regret, i.e., the regret under that policy is $\mathcal{O}(\log I)$.

Let $\gamma \in \Gamma^{(0)} \subset \mathbb{R}$ denote the unknown parameter of the win curve $p(\gamma, b)$. Recall that $Q^{b,\gamma}$ is the probability distribution of the outcome $D$ of the winning of impression for a given bid $b$. Under the well-separated condition, the win curve satisfies the following two assumptions.

**Assumption 5** *For all bids $b \in B$,*

1. *The family of distributions $\{Q^{b,\gamma} : \gamma \in \Gamma^{(0)}\}$ is identifiable.*
2. *There exists a constant $c_f > 0$ such that the Fisher information $I(b,\gamma)$, given by*

$$I(b,\gamma) = \mathbb{E}\left[-\frac{\partial^2}{\partial\gamma^2}\log Q^{b,\gamma}(D)\right]$$

*satisfies $I(b,\gamma) \geq c_f$ for all $\gamma \in \Gamma^{(0)}$.*

**Assumption 6** *For any sequence of bids $\boldsymbol{b} = (b_1, \cdots, b_k) \in B^k$ and $\boldsymbol{d} \in \{0,1\}^k$, $-\log Q^{\boldsymbol{b},\gamma}(\boldsymbol{d})$ is convex in $\gamma$ for $\gamma \in \Gamma^{(0)}$.*

### H.1 Lower Bound on the Regret

In this section, we derive an $\Omega(\log I)$ lower bound on the regret under any policy.

**Theorem A.3** *Consider the following instance of problem $(P)$: $B = [5/8, 7/8]$, $\Gamma^{(0)} = [2,3]$, $e = 1$, $T = 1$, and $I \geq 2$. There is only one location, i.e., $L = 1$. In each period, an impression arrives with probability $q > 0$. The probability of winning an arriving impression under a bid price $b \in B = [b^{\min}, b^{\max}]$ is $p(\gamma, b) = -1/2 + (b\gamma)/2$. There is one campaign in each season and the required number of impressions is no less than the number of periods in the season, so that the target of the campaign can never be exceeded. Then, for any policy $\pi$ setting bids in $B$, there exist $\gamma \in \Gamma^{(0)}$ and a constant $K_{13} > 0$ that is independent of $I$, such that*

$$\textit{Regret}(\pi, I; \gamma) \geq K_{13}\log(I).$$

Let $r(b,\gamma) = (b-1)[-1/2 + (b\gamma)/2]$. Then, $b^*(\gamma) = \frac{1}{2} + \frac{1}{2\gamma}$ minimizes $r(b,\gamma)$. Lemma A.12 below is used in the proof of Theorem A.3.

**Lemma A.12** *Consider the problem instance defined in the statement of Theorem 5 above. Let $\hat{\gamma}$ be a random variable taking values in $\Gamma^{(0)} = [2,3]$, with density $\rho : \Gamma^{(0)} \to \mathbb{R}_+$ given by $\rho(\gamma) = 2\{\cos(\pi(\gamma - 5/2))\}^2$. Then, for any bidding policy $\pi$ and any season $i \geq 1$, there exists a constant $K_{12} > 0$ such that*

$$\mathbb{E}\left[(b^*(\hat{\gamma}) - b_{i+1})^2\right] \geq K_{12} \cdot \frac{1}{i},$$

*where $b_{i+1}$ is the bid placed by $\pi$ at season $i+1$, and $\mathbb{E}[\cdot]$ denotes the expectation with respect to the joint distribution of $b_i$ and the prior density $\rho$ of the parameter $\hat{\gamma} \in \Gamma^{(0)}$.*

The proof of Lemma A.12 is similar to the proof of Lemma 4.6 in Broder and Rusmevichientong (2012), and thus is omitted for brevity.

**Proof of Theorem A.3:** It is straightforward to check that

$$r(b,\gamma) - r(b^*(\gamma),\gamma) \geq (b^*(\gamma) - b)^2.$$

Then, we have

$$\begin{aligned}
&\sup_{\gamma \in \Gamma^{(0)}} \text{Regret}(\pi, I; \gamma) \\
&\geq q \sup_{\gamma \in \Gamma^{(0)}} \sum_{i=1}^{I-1} \mathbb{E}[r(b_{i+1},\gamma) - r(b^*(\gamma),\gamma)] \\
&\geq q \sum_{i=1}^{I-1} \mathbb{E}[r(b_{i+1},\hat{\gamma}) - r(b^*(\hat{\gamma}),\hat{\gamma})] \\
&\geq q \sum_{i=1}^{I-1} \mathbb{E}[(b^*(\hat{\gamma}) - b_{i+1})^2] \\
&\geq qK_{12} \sum_{i=1}^{I-1} \frac{1}{i} \\
&\geq K_{13} \log(I),
\end{aligned}$$

where $K_{13} = qK_{12}$. The fourth inequality holds by Lemma A.12. ■

## H.2 Upper Bound on the Regret

In this section, we present a bidding policy similar to the pricing policy presented in Broder and Rusmevichientong (2012), and show that the regret under our policy is $\mathcal{O}(\log I)$. Let $\tilde{\pi}$ denote the bidding policy defined below.

**Inputs:** An initial bid $b_1 \in B$.

**Initialization:** When the first impression arrives, place the bid $b_1$ and observe the corresponding outcome $D_1$.

**Description:** For $\tau \geq 2$, when the $\tau^{th}$ impression arrives:

- Compute the maximum-likelihood estimate $\hat{\gamma}(\tau-1)$ given by

$$\hat{\gamma}(\tau-1) = \arg\max_{\gamma\in\Gamma^{(0)}} Q^{\tilde{\pi},\gamma}(\boldsymbol{D}_{\tau-1}).$$

  where $\boldsymbol{D}_{\tau-1} = (D_1, \cdots D_{\tau-1})$ denotes the observed outcome under policy $\tilde{\pi}$ for the first $\tau-1$ arrived impressions and

$$Q^{\pi,\gamma}(\boldsymbol{D}_{\tau-1}) = \prod_{\hat{\tau}=1}^{\tau-1} \left[p(b_{\hat{\tau}}^{\pi}(\boldsymbol{D}_{\hat{\boldsymbol{\tau}}-\mathbf{1}}),\gamma)^{D_{\hat{\tau}}}(1-p(b_{\hat{\tau}}^{\pi}(\boldsymbol{D}_{\hat{\boldsymbol{\tau}}-\mathbf{1}}),\gamma))^{1-D_{\hat{\tau}}}\right]$$

  is the probability of observing the realization $\boldsymbol{D}_{\boldsymbol{\tau}-\mathbf{1}}$ under policy $\pi$ when the underlying parameter is $\gamma$.

- Let $(i(\tau), t(\tau))$ denote the period when the $\tau^{th}$ impression arrives. Place bid $b^D_{i(\tau),t(\tau)}(\boldsymbol{c}+w\boldsymbol{e}_{\boldsymbol{\tau}};\hat{\gamma}(\tau-1))$ (see Section 4) based on the estimate $\hat{\gamma}(\tau-1)$.

Next, we show the regret under policy $\tilde{\pi}$ is $\mathcal{O}(\log I)$.

**Theorem A.4** *Under Assumptions 1, 4, 5, and 6, For any initial bid $b_1 \in B$, $T \geq 3$ and $I \geq 3$, policy $\tilde{\pi}$ satisfies*

$$\textit{Regret}(\tilde{\pi}, I; \gamma^{(0)}) \leq K_{17}\log I,$$

*where $K_{17} = K_{16}(K_5)^{2T}\log(T)$ for constants $K_5$ and $K_{16}$ that are independent of $I$ and $T$.*

Lemma A.13 below is used in the proof of Theorem A.4.

**Lemma A.13 (Theorem 4.7 in Broder and Rusmevichientong 2012)** *Let $\hat{\gamma}(\tau)$ be the maximum-likelihood estimate based on the observed outcomes for the first $\tau$ arrived impressions. Under Assumptions 1, 4, 5, and 6, there exists a constant $c_H$ such that for any $\tau \geq 1$, $\gamma \in \Gamma^{(0)}$, and $\epsilon \geq 0$,*

$$\Pr\{|\hat{\gamma}(\tau)-\gamma^{(0)}| \geq \epsilon\} \leq 2\exp(-\tau c_H\epsilon^2/2) \text{ and } \mathbb{E}[|\hat{\gamma}(\tau)-\gamma^{(0)}|^2] \leq \frac{4}{c_H}\cdot\frac{1}{\tau}.$$

**Proof of Theorem A.4:**

$$\begin{aligned}
&\text{Regret}\left(\tilde{\pi}, I; \gamma^{(0)}\right)\\
&\leq K_0\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T} q\left(b_{i,t}^{\tilde{\pi},D}(X_{i,1}^{\tilde{\pi},D}, \hat{X}_{i,t}^{\tilde{\pi},D}, W_{i,t}, \mathcal{T}_{i,t}) - b_{i,t}^{D}\left(\boldsymbol{c}(\hat{X}_{i,t}^{\tilde{\pi},D}) + W_{i,t}\boldsymbol{e}_{\boldsymbol{\mathcal{T}_{i,t}}};\gamma^{(0)}\right)\right)^2\right]\\
&\leq K_0(b^{\max}-b^{\min})^2 + K_0\mathbb{E}\left[\sum_{\tau=2}^{\mathcal{T}}\left(b_{i,t}^{D}(c(\hat{X}_{i,t}^{\tilde{\pi},D}) + W_{i,t}\boldsymbol{e}_{\boldsymbol{\mathcal{T}_{i,t}}};\hat{\gamma}(\tau-1)) - b_{i,t}^{D}\left(c(\hat{X}_{i,t}^{\tilde{\pi},D}) + W_{i,t}\boldsymbol{e}_{\boldsymbol{\mathcal{T}_{i,t}}};\gamma^{(0)}\right)\right)^2\right]\\
&\leq K_0(b^{\max}-b^{\min})^2 + K_0(K_5)^{2T}\mathbb{E}\left[\sum_{\tau=1}^{\mathcal{T}-1}(\hat{\gamma}(\tau)-\gamma^{(0)})^2\right]\\
&\leq K_0(b^{\max}-b^{\min})^2 + K_0(K_5)^{2T}\frac{4}{c_H}\sum_{\tau=1}^{IT}\frac{1}{\tau}\\
&\leq K_0(b^{\max}-b^{\min})^2 + K_0(K_5)^{2T}\frac{4}{c_H}[1+\log(IT)]
\end{aligned}$$

$$\leq K_{14} + K_{15}(K_5)^{2T}\log(T)\log(I)$$

$$\leq K_{17}\log(I),$$

where $K_{14} = K_0(b^{\max} - b^{\min})^2$, $K_{15} = 16K_0/c_H$, and $K_{17} = (K_{14} + K_{15})(K_5)^{2T}\log(T)$. The first inequality holds by Lemma 5. The third inequality holds by Lemma 4. The fourth inequality holds by Lemma A.13. Let $K_{16} = K_{14} + K_{15}$. This completes the proof of Theorem A.4. ■

## Appendix I Analysis of the Regret Under a Given Allocation Policy

In this section, we consider a setting where we are *given* an arbitrary and non-anticipating (deterministic) allocation policy in the set $\Phi$ (defined in Remark 5 of the main paper). Specifically, the active campaign to which an impression acquired in period $(i,t)$ from location $l$ is assigned (i.e., the allocation decision $a_{i,t,l}$) is deterministically defined based on the history in season $i$ until the beginning of that period, i.e., $\hat{x}_{i,t}$. Examples of such allocation policies include the FEFS policy and the policy that allocates an acquired impression to the active campaign with the highest ratio of penalty cost to the remaining duration of the campaign. Given any such allocation policy, the platform only needs to determine its bidding policy.

For convenience, we now recall the formulation of the platform's problem where the penalty costs are different across campaigns and the given allocation policy belongs to the set $\Phi$. Let $e_{i,j} \in [e^{\min}, e^{\max}]$ denote the unit penalty cost of campaign $(i,j)$. The total expected cost (i.e., the bidding cost plus the penalty cost) under policy $\pi$ after $I$ seasons is

$$\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t} + \sum_{j=1}^{m_i}\left[W_{i,j} - \sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a_{i,t,l}=j\}\right]^{+} e_{i,j}\right]. \tag{A-24}$$

Recall that, for any campaign $(i,j)$, it is optimal for the platform to not assign more than $W_{i,j}$ impressions to that campaign. We let $\Pi$ denote the set of all non-anticipating policies satisfying $\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t} = l\}d_{i,t}\mathbb{1}\{a_{i,t,l}=j\} \leq W_{i,j}$ a.s. for all $(i,j)\in\mathcal{C}_I$. Under any policy $\pi\in\Pi$, we can rewrite the total expected cost in (A-24) as

$$\begin{aligned}
&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t} + \sum_{j=1}^{m_i}\left[W_{i,j} - \sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\mathbb{1}\{a_{i,t,l}=j\}\right] e_{i,j}\right]\\
=&\sum_{i=1}^{I}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} b^{\pi}_{i,t,l}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t} - \sum_{t=1}^{T}\sum_{l\in\mathcal{L}}\mathbb{1}\{\zeta_{i,t}=l\}d_{i,t}\sum_{j=1}^{m_i} e_{i,j}\mathbb{1}\{a_{i,t,l}=j\}\right] + \sum_{i=1}^{I}\sum_{j=1}^{m_i} e_{i,j}W_{i,j}\\
=&\,\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b^{\pi}_{i,t,l} - e_{i,a_{i,t,l}})q_l d_{i,t}\right] + \sum_{i=1}^{I}\sum_{j=1}^{m_i} e_{i,j}W_{i,j}.
\end{aligned}$$

The second equality holds, since $\mathbb{E}[\mathbb{1}\{\zeta_{i,t} = l\}] = q_l$. Thus, the platform's problem can now be equivalently written as

$$\min_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}}(b^{\pi}_{i,t,l} - e_{i,a_{i,t,l}})q_l d_{i,t}\right].$$

We consider the clairvoyant problem in season $i$ and formulate a DP that defines its optimal cost. For a given allocation policy $\phi$ in $\Phi$, let $a^{\phi}_{i,t,l}(\hat{x}_{i,t})$ denote the active campaign to which an impression acquired in period $(i,t)$ from location $l$ is assigned. For ease of exposition, we henceforth drop the superscript $\phi$ in $a^{\phi}_{i,t,l}(\hat{x}_{i,t})$ when there is no ambiguity. Recall that $c(\hat{x}_{i,t})$ is the associated total number of unmet impressions at the beginning of period $(i,t)$ over all the ongoing campaigns in that period. Let $V_{i,t}(\hat{x}_{i,t};\gamma)$ denote the optimal cost-to-go function of the DP and let $b^{*}_{i,t,l}(\hat{x}_{i,t};\gamma)$ denote the optimal bid price at location $l$ in period $(i,t)$ in state $\hat{x}_{i,t}$. Then, $V_{i,t}(\hat{x}_{i,t};\gamma)$ satisfies the following recursion:

$$
\begin{aligned}
&V_{i,t}(\hat{x}_{i,t};\gamma)\\
&=\min_{\substack{(b_1,\cdots,b_L):\\ b_l\in B_l, l\in\mathcal{L}}}\left\{\begin{array}{l}
\mathbb{1}\{c(\hat{x}_{i,t})\geq 1\}\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b_l)\left[b_l - e_{i,a_{i,t,l}(\hat{x}_{i,t})} + V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma)\right] \ + \\
\sum_{l\in\mathcal{L}} q_l\left[1-\mathbb{1}\{c(\hat{x}_{i,t})\geq 1\}p_l(\gamma_l,b_l)\right]V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) \ + \\
\left(1-\sum_{l\in\mathcal{L}} q_l\right)V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma)
\end{array}\right\}\\
&=\mathbb{1}\{c(\hat{x}_{i,t})\geq 1\}\sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\gamma_l,b_l)\left[b_l - e_{i,a_{i,t,l}(\hat{x}_{i,t})} - \Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)\right] +\\
&\sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) + \left(1-\sum_{l\in\mathcal{L}} q_l\right)V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma)
\end{aligned}
$$

and $V_{i,T+1}(\cdot;\gamma)=0$, where

$$\Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) = V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma).$$

For $c(\hat{x}_{i,t})\geq 1$, the optimal bid price when an impression arrives from location $l\in\mathcal{L}$ is as follows:

$$b^{*}_{i,t,l}(\hat{x}_{i,t};\gamma) = \arg\min_{b_l\in B} p_l(\gamma_l,b_l)\left[b_l - e_{i,a_{i,t,l}(\hat{x}_{i,t})} - \Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)\right].$$

**Lemma A.14** *For all $l\in\mathcal{L}$, $\gamma,\hat{\gamma}\in\Gamma^{(0)}$, $c(\hat{x}_{i,t})\geq 1$, $1\leq i\leq I$, and $1\leq t\leq T$, there exists a constant $K_{18}>0$ such that*

$$\left|b^{*}_{i,t,l}(\hat{x}_{i,t};\gamma) - b^{*}_{i,t,l}(\hat{x}_{i,t};\hat{\gamma})\right| \leq K_{18}\|\gamma-\hat{\gamma}\|.$$

The proof of Lemma A.14 is provided in Appendix J.

When $\gamma$ is unknown, conditional on $x_{i,1}$, the expected cost-to-go in season $i$ for state $\hat{x}_{i,t}$ under policy $\pi$, denoted by $V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)$, satisfies the following recursion:

$$
\begin{aligned}
&V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma) =\\
&\mathbb{1}\{c(\hat{x}_{i,t})\geq 1\}\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - e_{i,a_{i,t,l}(\hat{x}_{i,t})} + V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma)\right] \ +\\
&\sum_{l\in\mathcal{L}} q_l\left[1-\mathbb{1}\{c(\hat{x}_{i,t})\geq 1\}p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma) \ +
\end{aligned}
$$

$$\left(1-\sum_{l\in\mathcal{L}} q_l\right) V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma),$$

and $V^{\pi}_{i,T+1}(\cdot;x_{i,1},\gamma)=0$.

Under policy $\pi$, recall that $X^{\pi}_{i,1}$ is the random history until the beginning of season $i$ and $\hat{X}^{\pi}_{i,t}$ is the random history in season $i$ until the beginning of period $(i,t)$. Then, the expected cost under policy $\pi$ after $I$ seasons is $\sum_{i=1}^{I}\mathbb{E}\left[V^{\pi}_{i,1}(\emptyset;X^{\pi}_{i,1},\gamma^{(0)})\right]$ and the one under the optimal policy of the clairvoyant problem is $\sum_{i=1}^{I} V_{i,1}(\emptyset;\gamma^{(0)})$, and thus the regret under policy $\pi$ after $I$ seasons is:

$$\text{Regret}(\pi,I;\gamma^{(0)})=\sum_{i=1}^{I}\mathbb{E}\left[V^{\pi}_{i,1}(\emptyset;X^{\pi}_{i,1},\gamma^{(0)})\right]-\sum_{i=1}^{I} V_{i,1}(\emptyset;\gamma^{(0)}).$$

**Lemma A.15** *There exists a constant $K_{19}>0$ such that the regret under any bidding policy $\pi$ after $I$ seasons satisfies*

$$\textit{Regret}(\pi,I;\gamma^{(0)}) \le K_{19}\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,t,l}(X^{\pi}_{i,1},\hat{X}^{\pi}_{i,t};\gamma^{(0)})-b^{*}_{i,t,l}(\hat{X}^{\pi}_{i,t};\gamma^{(0)})\right)^2\right].$$

The proof of Lemma A.15 is provided in Appendix J.

Consider a new policy denoted by $\check{\pi}$, where the given allocation policy is an arbitrary policy in $\Phi$, and the bidding policy is similar to the one in BidAlloc except that the bidding amount in the exploitation phase of cycle $s$ is replaced by $b^{*}_{i,t,l}(\hat{x}_{i,t};\hat{\gamma}(s))$ based on the vector of estimates $\hat{\gamma}(s)$. Then, we have

**Theorem A.5** *Under Assumptions 1, 2, and 3, the policy $\check{\pi}$ satisfies*

$$\textit{Regret}(\check{\pi},I;\gamma^{(0)})\le K_{21}\sqrt{I},$$

*where $K_{21}=K_{20}K_{19}(K_{18})^2\sqrt{T}$ for constants $K_{18}$ and $K_{19}$ that are independent[10] of $I$, and $K_{20}$ that is independent of $I$ and $T$.*

The proof of Theorem A.5 is similar to that of Theorem 1, and thus is omitted for brevity.

## Appendix J Proofs of Lemmas A.14 and A.15

First, we show two lemmas that are used to show Lemmas A.14 and A.15.

Let $\tilde{x}_{i,t}$ and $\breve{x}_{i,t}$ denote two histories in season $i$ until the beginning of period $(i,t)$. For ease of exposition, we drop $\gamma$ in $V_{i,t}(\hat{x}_{i,t};\gamma)$ and $b^{*}_{i,t,l}(\hat{x}_{i,t};\gamma)$ in Lemma A.16 when there is no ambiguity.

[10]Here, the allocation policy is arbitrary. Therefore, the values of the constants $K_{19}$ and $K_{18}$ may depend on the given allocation policy, which may also depend on $T$. Since the allocation policy is arbitrary, it is difficult to isolate the exact dependence of $K_{19}$ and $K_{18}$ on $T$. When the allocation policy is FEFS, we have $K_{19}=K_0$ and $K_{18}=(K_1)^T$, where $K_0$ and $K_1$ are independent of $I$ and $T$.

**Lemma A.16** *For $i \in \{1, \cdots, I\}$, $2 \le t \le T+1$, and $\gamma \in \Gamma^{(0)}$, there exists $k_t \ge 0$ such that $|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \le k_t$. Let $K_{22} = \max_t k_t$. Then $|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \le K_{22}$.*

**Proof of Lemma A.16**: The proof is by induction on $t$. For $t = T+1$, $|V_{i,T+1}(\tilde{x}_{i,t}) - V_{i,T+1}(\check{x}_{i,t})| = 0$. For $2 \le t \le T$, suppose that $|V_{i,t+1}(\tilde{x}_{i,t+1}) - V_{i,t}(\check{x}_{i,t+1})| \le k_{t+1}$. We now show that $|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \le k_t$ using the following four cases. For simplicity of exposition, we drop the indices $i$ and $t$ of $a_{i,t,l}$ and $b^*_{i,t,l}$ below.

- Case 1: $c(\check{x}_{i,t}) \ge 1$ and $c(\tilde{x}_{i,t}) \ge 1$.

$$
\begin{aligned}
& V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t}) \\
= & \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l \in \mathcal{L}}} \left\{ \begin{array}{l} \sum\limits_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l) \left[ b_l - e_{i,a_l(\tilde{x}_{i,t})} + V_{i,t+1}((\tilde{x}_{i,t},(l,1))) \right] \; + \\ \sum\limits_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b_l)\right] V_{i,t+1}((\tilde{x}_{i,t},(l,0))) + \left(1 - \sum\limits_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\tilde{x}_{i,t},(0,0))) \end{array} \right\} - \\
& \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l \in \mathcal{L}}} \left\{ \begin{array}{l} \sum\limits_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l) \left[ b_l - e_{i,a_l(\check{x}_{i,t})} + V_{i,t+1}((\check{x}_{i,t},(l,1))) \right] \; + \\ \sum\limits_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b_l)\right] V_{i,t+1}((\check{x}_{i,t},(l,0))) + \left(1 - \sum\limits_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\check{x}_{i,t},(0,0))) \end{array} \right\} \\
\ge & \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\tilde{x}_{i,t}))[V_{i,t+1}((\tilde{x}_{i,t},(l,1))) - V_{i,t+1}((\check{x}_{i,t},(l,1)))] \; + e^{\min} - e^{\max} + \\
& \sum_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b^*_l(\tilde{x}_{i,t}))\right] [V_{i,t+1}((\tilde{x}_{i,t},(l,0))) - V_{i,t+1}((\check{x}_{i,t},(l,0)))] + \\
& \left(1 - \sum_{l\in\mathcal{L}} q_l\right) [V_{i,t+1}((\tilde{x}_{i,t},(0,0))) - V_{i,t+1}((\check{x}_{i,t},(0,0)))]. \qquad \text{(A-25)}
\end{aligned}
$$

The inequality holds by letting $b_l = b^*_l(\tilde{x}_{i,t})$.

$$
\begin{aligned}
& V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t}) \\
= & \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l \in \mathcal{L}}} \left\{ \begin{array}{l} \sum\limits_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l) \left[ b_l - e_{i,a_l(\tilde{x}_{i,t})} + V_{i,t+1}((\tilde{x}_{i,t},(l,1))) \right] \; + \\ \sum\limits_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b_l)\right] V_{i,t+1}((\tilde{x}_{i,t},(l,0))) + \left(1 - \sum\limits_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\tilde{x}_{i,t},(0,0))) \end{array} \right\} - \\
& \min_{\substack{(b_1,\cdots,b_L):\\ b_l \in B_l, l \in \mathcal{L}}} \left\{ \begin{array}{l} \sum\limits_{l\in\mathcal{L}} q_l p_l(\gamma_l, b_l) \left[ b_l - e_{i,a_l(\check{x}_{i,t})} + V_{i,t+1}((\check{x}_{i,t},(l,1))) \right] \; + \\ \sum\limits_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b_l)\right] V_{i,t+1}((\check{x}_{i,t},(l,0))) + \left(1 - \sum\limits_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\check{x}_{i,t},(0,0))) \end{array} \right\} \\
\le & \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\check{x}_{i,t}))[V_{i,t+1}((\tilde{x}_{i,t},(l,1))) - V_{i,t+1}((\check{x}_{i,t},(l,1)))] \; + e^{\max} - e^{\min} + \\
& \sum_{l\in\mathcal{L}} q_l \left[1 - p_l(\gamma_l, b^*_l(\check{x}_{i,t}))\right] [V_{i,t+1}((\tilde{x}_{i,t},(l,0))) - V_{i,t+1}((\check{x}_{i,t},(l,0)))] + \\
& \left(1 - \sum_{l\in\mathcal{L}} q_l\right) [V_{i,t+1}((\tilde{x}_{i,t},(0,0))) - V_{i,t+1}((\check{x}_{i,t},(0,0)))]. \qquad \text{(A-26)}
\end{aligned}
$$

The inequality holds by letting $b_l = b^*_l(\check{x}_{i,t})$. Combining (A-25) and (A-26), we have

$$|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \le k_{t+1} + e^{\max} - e^{\min}.$$

- Case 2: $c(\check{x}_{i,t}) = 0$ and $c(\tilde{x}_{i,t}) \geq 1$.

$$
\begin{aligned}
&|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \\
&= \left| \begin{array}{l} \sum_{l\in\mathcal{L}} q_l \min_{b_l\in B} p_l(\gamma_l, b_l)\left[b_l - e_{i,a_l(\tilde{x}_{i,t})} - \Delta V_{i,t+1}((\tilde{x}_{i,t},(l,0)))\right] + \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\tilde{x}_{i,t},(l,0)))+ \\ \left(1-\sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\tilde{x}_{i,t},(0,0))) - \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\check{x}_{i,t},(l,0))) - \left(1-\sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\check{x}_{i,t},(0,0))) \end{array} \right| \\
&\leq \max\{|b^{\min} - e^{\max}|, |b^{\max} - e^{\min}|\} + 2k_{t+1}.
\end{aligned}
$$

  The inequality holds by the induction hypothesis.

- Case 3: $c(\check{x}_{i,t}) \geq 1$ and $c(\tilde{x}_{i,t}) = 0$. Similar to the argument in Case 2, we have

$$|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \leq \max\{|b^{\min} - e^{\max}|, |b^{\max} - e^{\min}|\} + 2k_{t+1}.$$

- Case 4: $c(\check{x}_{i,t}) = 0$ and $c(\tilde{x}_{i,t}) = 0$.

$$
\begin{aligned}
&|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \\
&= \left| \begin{array}{l} \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\tilde{x}_{i,t},(l,0))) + \left(1-\sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\tilde{x}_{i,t},(0,0)))- \\ \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\check{x}_{i,t},(l,0))) - \left(1-\sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\check{x}_{i,t},(0,0))) \end{array} \right| \\
&\leq k_{t+1}.
\end{aligned}
$$

  The inequality holds by the induction hypothesis. Let $k_t = \max\{k_{t+1} + e^{\max} - e^{\min}, \max\{b^{\min} - e^{\max}|, |b^{\max} - e^{\min}|\} + 2k_{t+1}\}$. Then, we have $|V_{i,t}(\tilde{x}_{i,t}) - V_{i,t}(\check{x}_{i,t})| \leq k_t$ ■

**Lemma A.17** *For any $l \in \mathcal{L}$, $b_l \in B$, $\alpha \in \mathbb{R}$ and $\gamma_l \in \Gamma_l$, define the function*

$$f^l(b_l, \alpha, \gamma_l) = p_l(\gamma_l, b_l)(b_l - \alpha).$$

*Let $b_l^*(\alpha, \gamma_l) = \arg\min_{b_l\in B} f^l(b_l, \alpha, \gamma_l)$. Thus, for $c(\hat{x}_{i,t}) \geq 1$, we have $b_{i,t,l}^*(\hat{x}_{i,t}; \gamma) = b_l^*(\alpha, \gamma_l)$, where $\alpha = e_{i,a_{i,t,l}(\hat{x}_{i,t})} + \Delta V_{i,t+1}((\hat{x}_{i,t},(l,0)); \gamma)$. Then:*

*(i) For each $(\alpha, \gamma_l) \in \mathbb{R} \times \Gamma_l^{(0)}$, $b_l^*(\alpha, \gamma_l)$ is uniquely defined.*

*(ii) Let $\mathcal{U}_{A\Gamma_l} = \left\{(\alpha, \gamma_l) \in [-K_{22} + e^{\min}, K_{22} + e^{\max}] \times \Gamma_l^{(0)} \,\middle|\, b^{\min} < b_l^*(\alpha, \gamma_l) < b^{\max}\right\}$. For each $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$,*

$$\left.\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l}\right|_{b_l = b_l^*(\alpha,\gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2}\right|_{b_l = b_l^*(\alpha,\gamma_l)} > 0.$$

*(iii) For each $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$, both $b_l^*(\alpha, \gamma_l)$ and $f^l(b_l^*(\alpha, \gamma_l), \alpha, \gamma_l)$ are continuously differentiable in $\alpha$ and $\gamma_l$.*

*(iv) There exists a $K_{19} > 0$ such that $f^l(b_l, \alpha, \gamma_l) - f^l(b_l^*(\alpha, \gamma_l), \alpha, \gamma_l) \leq K_{19}(b_l - b_l^*(\alpha, \gamma_l))^2$ for all $b_l \in B$ and $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$.*

**Proof of Lemma A.17**:

**(i)** Let $(\alpha, \gamma_l) \in \mathbb{R} \times \Gamma_l^{(0)}$. We have

$$\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l} = p_l(\gamma_l, b_l) + (b_l - \alpha)\frac{\partial p_l(\gamma_l, b_l)}{\partial b_l}, \tag{A-27}$$

and

$$\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2} = 2\frac{\partial p_l(\gamma_l, b_l)}{\partial b_l} + (b_l - \alpha)\frac{\partial^2 p_l(\gamma_l, b_l)}{\partial b_l^2}.$$

It follows that any $b_l \in B$ with $\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l} = 0$ satisfies the following:

$$\begin{aligned}
\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2} &= 2\frac{\partial p_l(\gamma_l, b_l)}{\partial b_l} + \frac{-p_l(\gamma_l, b_l)}{\partial p_l(\gamma_l, b_l)/\partial b_l}\frac{\partial^2 p_l(\gamma_l, b_l)}{\partial b_l^2} \\
&= \frac{\partial p_l(\gamma_l, b_l)}{\partial b_l}\left[2 - \frac{p_l(\gamma_l, b_l)\partial^2 p_l(\gamma_l, b_l)/\partial b_l^2}{(\partial p_l(\gamma_l, b_l)/\partial b_l)^2}\right] \\
&= \frac{\partial p_l(\gamma_l, b_l)}{\partial b_l}\left[1 + \frac{(\partial p_l(\gamma_l, b_l)/\partial b_l)^2 - p_l(\gamma_l, b_l)\partial^2 p_l(\gamma_l, b_l)/\partial b_l^2}{p_l(\gamma_l, b_l)^2}\frac{p_l(\gamma_l, b_l)^2}{(\partial p_l(\gamma_l, b_l)/\partial b_l)^2}\right] \\
&= \frac{\partial p_l(\gamma_l, b_l)}{\partial b_l}\left[1 - \frac{\partial^2 \log(p_l(\gamma_l, b_l))}{\partial b_l^2}\frac{p_l(\gamma_l, b_l)^2}{(\partial p_l(\gamma_l, b_l)/\partial b_l)^2}\right] \\
&> 0.
\end{aligned}$$

The inequality holds since $\frac{\partial p_l(\gamma_l, b_l)}{\partial b_l} > 0$ by Assumption 1 (Section 2.1) and $\frac{\partial^2 \log(p_l(\gamma_l, b_l))}{\partial b_l^2} \leq 0$ by the log-concavity of $p_l(\gamma_l, b_l)$ with respect to $b_l$.

Thus, $f^l(b_l, \alpha, \gamma_l)$ either has a unique minimum $b_l^*(\alpha, \gamma_l) \in (b^{\min}, b^{\max})$ with

$$\left.\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l}\right|_{b_l = b_l^*(\alpha, \gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2}\right|_{b_l = b_l^*(\alpha, \gamma_l)} > 0,$$

or is monotone on $B$ and the unique minimum of $f^l(b_l, \alpha, \gamma_l)$ is on the boundary of $B$.

**(ii)** For $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$, since $b_l^*(\alpha, \gamma_l) \in (b^{\min}, b^{\max})$, we have

$$\left.\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l}\right|_{b_l = b_l^*(\alpha, \gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2}\right|_{b_l = b_l^*(\alpha, \gamma_l)} > 0.$$

**(iii)** We first show that $b_l^*(\alpha, \gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$ using the Implicit Function Theorem (see, e.g., Theorem 9.2 in Munkres 2018). Notice that

- By Assumption 1, $\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l}$ in Equation (A-27) is continuously differentiable in $\alpha$, $\gamma_l$, and $b_l$, on the open set $\mathbb{R} \times \Gamma_l \times (b^{\min}, b^{\max})$ .
- By (ii), for each $(\alpha, \gamma_l) \in \mathcal{U}_{A\Gamma_l}$, $(\alpha, \gamma_l, b_l^*(\alpha, \gamma_l))$ is a point in $\mathbb{R} \times \Gamma_l \times (b^{\min}, b^{\max})$ such that

$$\left.\frac{\partial f^l(b_l, \alpha, \gamma_l)}{\partial b_l}\right|_{b_l = b_l^*(\alpha, \gamma_l)} = 0 \text{ and } \left.\frac{\partial^2 f^l(b_l, \alpha, \gamma_l)}{\partial b_l^2}\right|_{b_l = b_l^*(\alpha, \gamma_l)} > 0.$$

Therefore, by the Implicit Function Theorem, $b_l^*(\alpha, \gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$.

Next, we show that $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$. The partial derivatives of $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ with respect to $\alpha$ and $\gamma_l$ are:

$$\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial \alpha} = \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} \frac{\partial b_l^*(\alpha,\gamma_l)}{\partial \alpha} + \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial \alpha}\right|_{b_l=b_l^*(\alpha,\gamma_l)} = -p_l(\gamma_l,b_l^*(\alpha,\gamma_l)),$$

$$\begin{aligned}\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial \gamma_l} &= \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} \frac{\partial b_l^*(\alpha,\gamma_l)}{\partial \gamma_l} + \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial \gamma_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} \\ &= (b_l^*(\alpha,\gamma_l)-\alpha)\left.\frac{\partial p_l(\gamma_l,b_l)}{\partial \gamma_l}\right|_{b_l^*(\alpha,\gamma_l)}.\end{aligned}$$

By Assumption 1 and the fact that $b_l^*(\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$, the above expressions of $\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial \alpha}$ and $\frac{\partial f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)}{\partial \gamma_l}$ are continuous in $\alpha$ and $\gamma_l$. Thus, $f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l)$ is continuously differentiable in $\alpha$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$.

**(iv)** Let $K_{19}^l := \sup_{(\alpha,\gamma_l,b_l)\in\mathcal{U}_{A\Gamma_l}\times B} \frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}/2$. Since $\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}$ is continuous in $\alpha$, $\gamma_l$, and $b_l$, on the closure of $\mathcal{U}_{A\Gamma_l}\times B$, which is compact, and $\left.\frac{\partial^2 f^l(b_l,\alpha,\gamma_l)}{\partial b_l^2}\right|_{b_l=b_l^*(\alpha,\gamma_l)} > 0$ for all $(\alpha,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$, we have $0 < K_{19}^l < \infty$. The Taylor expansion of $f^l(b_l,\alpha,\gamma_l)$ at $b_l = b_l^*(\alpha,\gamma_l)$ implies that

$$\begin{aligned}f^l(b_l,\alpha,\gamma_l) &\le f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) + \left.\frac{\partial f^l(b_l,\alpha,\gamma_l)}{\partial b_l}\right|_{b_l=b_l^*(\alpha,\gamma_l)} (b_l - b_l^*(\alpha,\gamma_l)) + K_{19}^l(b_l-b_l^*(\alpha,\gamma_l))^2 \\ &= f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) + K_{19}^l(b_l-b_l^*(\alpha,\gamma_l))^2.\end{aligned}$$

Let $K_{19} := \max_{l\in\mathcal{L}} K_{19}^l$. Then, we have $f^l(b_l,\alpha,\gamma_l) - f^l(b_l^*(\alpha,\gamma_l),\alpha,\gamma_l) \le K_{19}(b_l - b_l^*(\alpha,\gamma_l))^2$ for all $l\in\mathcal{L}$. ■

**Proof of Lemma A.14:** In the notation of Lemma A.17, let $\alpha_l = e_{i,a_{i,t,l}(\hat{x}_{i,t})} + \Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)$ and $\hat{\alpha}_l = e_{i,a_{i,t,l}(\hat{x}_{i,t})} + \Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\hat{\gamma})$. Then, $b_{i,t,l}^*(\hat{x}_{i,t};\gamma) - b_{i,t,l}^*(\hat{x}_{i,t};\hat{\gamma}) = b_l^*(\alpha_l,\gamma_l) - b_l^*(\hat{\alpha}_l,\hat{\gamma}_l)$. By Lemma A.16 and Assumption 3, $(\alpha_l,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$ and $(\hat{\alpha}_l,\hat{\gamma}_l)\in\mathcal{U}_{A\Gamma_l}$. Since $b_l^*(\alpha_l,\gamma_l)$ is continuously differentiable in $\alpha_l$ and $\gamma_l$ on $\mathcal{U}_{A\Gamma_l}$ by part (iii) of Lemma A.17 and by the fact that the closure of $\mathcal{U}_{A\Gamma_l}$ is compact, it follows from the first-order Taylor expansion that

$$|b_l^*(\alpha_l,\gamma_l) - b_l^*(\hat{\alpha}_l,\hat{\gamma}_l)| \le K_{23}^l(|\alpha_l-\hat{\alpha}_l| + \|\gamma_l-\hat{\gamma}_l\|), \tag{A-28}$$

for $K_{23}^l > 0$ that is independent of $\alpha_l$, $\hat{\alpha}_l$, $\gamma_l$, and $\hat{\gamma}_l$. We show by backward induction below (under the title "Derivation of Inequality (A-29)") that there exists $\kappa_t > 0$ such that

$$|V_{i,t}(\hat{x}_{i,t};\gamma) - V_{i,t}(\hat{x}_{i,t};\hat{\gamma})| \le \kappa_t\|\gamma-\hat{\gamma}\|. \tag{A-29}$$

Combining (A-28) and (A-29), we have

$$\begin{aligned}&|b_l^*(\alpha_l,\gamma_l) - b_l^*(\hat{\alpha}_l,\hat{\gamma}_l)| \\ \le& K_{23}^l(|\alpha_l-\hat{\alpha}_l| + \|\gamma_l-\hat{\gamma}_l\|) \\ \le& K_{23}^l(|V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,0));\hat{\gamma})| + \\ &|V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,1));\hat{\gamma})| + \|\gamma_l-\hat{\gamma}_l\|)\end{aligned}$$

$$\leq K_{23}^l \left( 2 \max_{t\in\{1,\ldots,T\}} \kappa_{t+1} \|\gamma - \hat{\gamma}\| + \|\gamma - \hat{\gamma}\| \right)$$
$$\leq K_{18} \|\gamma - \hat{\gamma}\|,$$

where $K_{18} = \max_{l\in\mathcal{L}} K_{23}^l \left(2 \max_{t\in\{1,\ldots,T\}} \kappa_{t+1} + 1\right)$. ■

**Derivation of Inequality** (A-29)**:** We show inequality (A-29) by backward induction on $t$. If $t = T+1$, then $V_{i,T+1}(\hat{x}_{i,T+1};\gamma) = V_{i,T+1}(\hat{x}_{i,T+1};\hat{\gamma}) = 0$ and (A-29) holds. Let $1 \leq t \leq T$. Suppose (A-29) holds for $t+1$. We now show that (A-29) holds for $t$.

$$\begin{aligned}
&|V_{i,t}(\hat{x}_{i,t};\gamma) - V_{i,t}(\hat{x}_{i,t};\hat{\gamma})| \\
&= \left| \begin{array}{l} \mathbb{1}\{c(\hat{x}_{i,t}) \geq 1\} \sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\gamma_l, b_l)(b_l - \alpha_l) + \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) + \\ \left(1 - \sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma) - \mathbb{1}\{c(\hat{x}_{i,t}) \geq 1\} \sum_{l\in\mathcal{L}} q_l \min_{b_l\in B_l} p_l(\hat{\gamma}_l, b_l)(b_l - \hat{\alpha}_l) - \\ \sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\hat{x}_{i,t},(l,0));\hat{\gamma}) - \left(1 - \sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\hat{x}_{i,t},(0,0));\hat{\gamma}) \end{array} \right| \\
&\leq \left| \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha_l,\gamma_l),\alpha_l,\gamma_l) - \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\hat{\alpha}_l,\hat{\gamma}_l),\hat{\alpha}_l,\hat{\gamma}_l) \right| + \kappa_{t+1}\|\gamma - \hat{\gamma}\| \\
&\leq K_{24}(|\alpha_l - \hat{\alpha}_l| + \|\gamma - \hat{\gamma}\|) + \kappa_{t+1}\|\gamma - \hat{\gamma}\| \\
&\leq K_{24}[|V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,0));\hat{\gamma})| + |V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,1));\hat{\gamma})|] + \\
&\quad (K_{24} + \kappa_{t+1})\|\gamma - \hat{\gamma}\| \\
&\leq K_{24}\left[\kappa_{t+1}\|\gamma - \hat{\gamma}\| + \kappa_{t+1}\|\gamma - \hat{\gamma}\|\right] + (K_{24} + \kappa_{t+1})\|\gamma - \hat{\gamma}\| \\
&= \kappa_t \|\gamma - \hat{\gamma}\|,
\end{aligned}$$

where $\kappa_t = 2K_{24}\kappa_{t+1} + K_{24} + \kappa_{t+1}$. The second inequality holds since $f^l(b_l^*(\alpha_l,\gamma_l),\alpha_l,\gamma_l)$ is continuously differentiable in $\alpha_l$ and $\gamma_l$, by part (iii) of Lemma A.17. Therefore, $\sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha_l,\gamma_l),\alpha_l,\gamma_l)$ is continuously differentiable in $\alpha_l$ and $\gamma$. In addition, $[-K_{22} + e^{\min}, K_{22} + e^{\max}] \times \Gamma^{(0)}$ is compact. It follows by a first-order Taylor expansion that there exists $K_{24} > 0$ that is independent of $\alpha_l$, $\hat{\alpha}_l$, $\gamma$, and $\hat{\gamma}$, such that

$$\left| \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\alpha_l,\gamma_l),\alpha_l,\gamma_l) - \sum_{l\in\mathcal{L}} q_l f^l(b_l^*(\hat{\alpha}_l,\hat{\gamma}_l),\hat{\alpha}_l,\hat{\gamma}_l) \right| \leq K_{24}(|\alpha_l - \hat{\alpha}_l| + \|\gamma - \hat{\gamma}\|).$$ ■

**Proof of Lemma A.15:** We establish the result by showing that there exists a constant $K_{19} > 0$ such that

$$\begin{aligned}
&V_{i,t}^{\pi}(\hat{x}_{i,t};x_{i,1},\gamma) - V_{i,t}(\hat{x}_{i,t};\gamma) \\
\leq &K_{19}\mathbb{E}\left[ \sum_{\hat{t}=t}^{T} \sum_{l\in\mathcal{L}} q_l \left( b_{i,\hat{t},l}^{\pi}\left(x_{i,1}, \hat{X}_{i,\hat{t}}^{\pi}\right) - b_{i,\hat{t},l}^{*}\left(\hat{X}_{i,\hat{t}}^{\pi};\gamma\right)\right)^2 \middle| \hat{X}_{i,t}^{\pi} = \hat{x}_{i,t} \right].
\end{aligned} \tag{A-30}$$

Then, the regret under any policy $\pi$ after $I$ seasons satisfies

$$\sum_{i=1}^{I} \mathbb{E}\left[ V_{i,1}^{\pi}(\emptyset; X_{i,1}^{\pi}, \gamma^{(0)}) - V_{i,1}(\emptyset;\gamma^{(0)}) \right]$$

$$\leq K_{19}\mathbb{E}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,t,l}\left(X^{\pi}_{i,1},\hat{X}^{\pi}_{i,t}\right)-b^{*}_{i,t,l}\left(\hat{X}^{\pi}_{i,t};\gamma^{(0)}\right)\right)^2\right].$$

Next, we show (A-30) using backward induction on $t$. For $t=T+1$, we have

$$V^{\pi}_{i,T+1}(\hat{x}_{i,T+1};x_{i,1},\gamma)=V_{i,T+1}(\hat{x}_{i,T+1};\gamma)=0.$$

Let $1\leq t\leq T$. Suppose (A-30) holds at $t+1$, i.e.,

$$\begin{aligned}&V^{\pi}_{i,t+1}(\hat{x}_{i,t+1};x_{i,1},\gamma)-V_{i,t+1}(\hat{x}_{i,t+1};\gamma)\\ &\leq K_{19}\mathbb{E}\left[\sum_{\hat{t}=t+1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}\left(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}}\right)-b^{*}_{i,\hat{t},l}\left(\hat{X}^{\pi}_{i,\hat{t}};\gamma\right)\right)^2\middle|\hat{X}^{\pi}_{i,t+1}=\hat{x}_{i,t+1}\right].\end{aligned}$$

Let $\alpha_l=e_{i,a_{i,t,l}(\hat{x}_{i,t})}+\Delta V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)$. Then, $b^{*}_{i,t,l}(\hat{x}_{i,t};\gamma)=b^{*}_{l}(\alpha_l,\gamma_l)$. By Lemma A.16 and Assumption 3, we have $(\alpha_l,\gamma_l)\in\mathcal{U}_{A\Gamma_l}$.

- Case 1: If $c(\hat{x}_{i,t})=0$, then we have

$$\begin{aligned}&V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)-V_{i,t}(\hat{x}_{i,t};\gamma)\\ &=\sum_{l\in\mathcal{L}} q_l V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma)+\left(1-\sum_{l\in\mathcal{L}} q_l\right)V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma)-\\ &\quad\sum_{l\in\mathcal{L}} q_l V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)-\left(1-\sum_{l\in\mathcal{L}} q_l\right)V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma)\\ &=\mathbb{E}\left[V^{\pi}_{i,t+1}(\hat{X}^{\pi}_{i,t+1};x_{i,1},\gamma)-V_{i,t+1}(\hat{X}^{\pi}_{i,t+1};\gamma)\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right]\\ &\leq\mathbb{E}\left[K_{19}\mathbb{E}\left[\sum_{\hat{t}=t+1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}})-b^{*}_{i,\hat{t},l}(\hat{X}^{\pi}_{i,\hat{t}};\gamma)\right)^2\middle|\hat{X}^{\pi}_{i,t+1}\right]\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right]\\ &=K_{19}\mathbb{E}\left[\sum_{\hat{t}=t}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^{\pi}_{i,\hat{t},l}(x_{i,1},\hat{X}^{\pi}_{i,\hat{t}})-b^{*}_{i,\hat{t},l}(\hat{X}^{\pi}_{i,\hat{t}};\gamma)\right)^2\middle|\hat{X}^{\pi}_{i,t}=\hat{x}_{i,t}\right]\end{aligned}$$

  The inequality holds by the induction hypothesis.

- Case 2: If $c(\hat{x}_{i,t})\geq 1$, then we have

$$\begin{aligned}&V^{\pi}_{i,t}(\hat{x}_{i,t};x_{i,1},\gamma)-V_{i,t}(\hat{x}_{i,t};\gamma)\\ &=\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t})-e_{i,a_{i,t,l}(\hat{x}_{i,t})}+V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma)\right]+\\ &\sum_{l\in\mathcal{L}} q_l\left[1-p_l(\gamma_l,b^{\pi}_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]V^{\pi}_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma)+\\ &\left(1-\sum_{l\in\mathcal{L}} q_l\right)V^{\pi}_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma)-\end{aligned}$$

$$
\begin{aligned}
&\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_{i,t,l}(\hat{x}_{i,t};\gamma))\left[b^*_{i,t,l}(\hat{x}_{i,t};\gamma) - e_{i,a_{i,t,l}(\hat{x}_{i,t})} + V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma)\right] \; - \\
&\sum_{l\in\mathcal{L}} q_l\left[1 - p_l(\gamma_l, b^*_{i,t,l}(\hat{x}_{i,t};\gamma))\right] V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) \; - \\
&\left(1 - \sum_{l\in\mathcal{L}} q_l\right) V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma) \\
= &\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[V^\pi_{i,t+1}((\hat{x}_{i,t},(l,1));x_{i,1},\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma)\right] \; + \\
&\sum_{l\in\mathcal{L}} q_l\left[1 - p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right]\left[V^\pi_{i,t+1}((\hat{x}_{i,t},(l,0));x_{i,1},\gamma) - V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma)\right] \; + \\
&\left(1 - \sum_{l\in\mathcal{L}} q_l\right)\left[V^\pi_{i,t+1}((\hat{x}_{i,t},(0,0));x_{i,1},\gamma) - V_{i,t+1}((\hat{x}_{i,t},(0,0));\gamma)\right] \; + \\
&\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma) + \\
&\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - e_{i,a_{i,t,l}(\hat{x}_{i,t})}\right] \; + \\
&\sum_{l\in\mathcal{L}} q_l\left[1 - p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\right] V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) \; - \\
&\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_{i,t,l}(\hat{x}_{i,t};\gamma))\left[b^*_{i,t,l}(\hat{x}_{i,t};\gamma) - e_{i,a_{i,t,l}(\hat{x}_{i,t})} + V_{i,t+1}((\hat{x}_{i,t},(l,1));\gamma)\right] \; - \\
&\sum_{l\in\mathcal{L}} q_l\left[1 - p_l(\gamma_l, b^*_{i,t,l}(\hat{x}_{i,t};\gamma))\right] V_{i,t+1}((\hat{x}_{i,t},(l,0));\gamma) \\
= &\,\mathbb{E}\left[V^\pi_{i,t+1}(\hat{X}^\pi_{i,t+1};x_{i,1},\gamma) - V_{i,t+1}(\hat{X}^\pi_{i,t+1};\gamma)\middle|\hat{X}^\pi_{i,t} = \hat{x}_{i,t}\right] \; + \\
&\sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}))\left[b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - \alpha_l\right] - \sum_{l\in\mathcal{L}} q_l p_l(\gamma_l, b^*_l(\alpha_l,\gamma_l))\left[b^*_l(\alpha_l,\gamma_l) - \alpha_l\right] \\
\le &\,\mathbb{E}\left[K_{19}\mathbb{E}\left[\sum_{\hat{t}=t+1}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^\pi_{i,\hat{t},l}(x_{i,1},\hat{X}^\pi_{i,\hat{t}}) - b^*_{i,\hat{t},l}(\hat{X}^\pi_{i,\hat{t}})\right)^2\middle|\hat{X}^\pi_{i,t+1}\right]\middle|\hat{X}^\pi_{i,t} = \hat{x}_{i,t}\right] \; + \\
&K_{19}\sum_{l\in\mathcal{L}} q_l\left(b^\pi_{i,t,l}(x_{i,1},\hat{x}_{i,t}) - b^*_l(\alpha_l,\gamma_l)\right)^2 \\
= &\,K_{19}\mathbb{E}\left[\sum_{\hat{t}=t}^{T}\sum_{l\in\mathcal{L}} q_l\left(b^\pi_{i,\hat{t},l}(x_{i,1},\hat{X}^\pi_{i,\hat{t}}) - b^*_{i,\hat{t},l}(\hat{X}^\pi_{i,\hat{t}};\gamma)\right)^2\middle|\hat{X}^\pi_{i,t} = \hat{x}_{i,t}\right]
\end{aligned}
$$

The inequality holds by the induction hypothesis and part (iv) of Lemma A.17. ■

## Appendix K Comparison with Respect to Broder and Rusmevichientong (2012) and den Boer and Zwart (2015)

Broder and Rusmevichientong (2012) study the dynamic pricing of a single product with unlimited inventory, where the seller needs to learn the unknown parameters of a general parametric demand distribution. They show that the regret under any pricing policy is $\Omega(\sqrt{N})$, where $N$ is the number of consumer arrivals, and propose a pricing policy that achieves a regret of $\mathcal{O}(\sqrt{N})$. den Boer and Zwart (2015) study the dynamic pricing of

multiple products, each of which has a finite inventory and is sold over a selling season of finite length. Only one product is sold in a season and all the products share the same unknown parametric demand distribution. They show that the regret under any pricing policy is $\Omega(\log I)$ and offer a policy that achieves a regret of $\mathcal{O}(\log^2(I))$, where $I$ is the number of seasons in the planning horizon of interest.

We now discuss the core features that differentiate the mobile-promotion platform's procurement problem from these two pricing problems:

- The notion of a "product" in our context is fundamentally different from that in the two papers above. An acquired impression cannot be stored in inventory and must be allocated to a campaign instantaneously. Further, impressions are substitutable in the sense that a campaign's demand can possibly be fulfilled by impressions acquired from different locations, with each location characterized by its own arrival probability and its own win curve. The platform's cost for acquiring different impressions could be different, not only for those arising at different locations but also for those from the same location.
- In both Broder and Rusmevichientong (2012) and den Boer and Zwart (2015), the supply side is "inactive", in the sense that the former assumes unlimited inventory and the latter assumes a fixed amount of inventory of each product. In contrast, in our problem, the supply of impressions is uncertain. Specifically, both the arrival of impressions as well as their acquisition by the platform (via real-time bidding on an ad-exchange) are uncertain.
- An impression won by the platform can be assigned to any one of multiple ongoing campaigns. This results in the need to *allocate* the impression to a campaign. Such an allocation decision is not needed in Broder and Rusmevichientong (2012) and den Boer and Zwart (2015), where for a given product and a given price, a consumer determines whether or not to buy the product.

Next, we contrast our policies for the bidding and allocation of impressions with the pricing policies in Broder and Rusmevichientong (2012) and den Boer and Zwart (2015).

- As in Broder and Rusmevichientong (2012), our bidding policy is also cyclic, with each cycle consisting of an exploration phase and an exploitation phase. However, within each of these phases, our policy is significantly different. We now briefly highlight these differences. First, note that, in contrast to Broder and Rusmevichientong (2012), we consider impression arrivals from multiple locations. During the exploration phase of a cycle, we offer "exploration bids" sequentially to estimate the underlying win-curve parameters for each location based on the observed realizations of the winning of impressions. If no impression arrives, then no bid is placed and, therefore, no observation is generated. For each location, we need a sufficient (location-specific) number of impression arrivals to place our exploration bids and obtain a good estimate of the underlying parameters. As a result, not only is the length of the exploration phase random, but also the number of observations of the winning of impressions. Consequently, for some locations with

high impression-arrival probabilities, we may have more observations than needed to compute a good estimate of its underlying parameters. All this is in contrast to Broder and Rusmevichientong (2012), where there is only one product and hence the number of observations (i.e., consumer arrivals) is fixed in each exploration phase.

During the exploitation phase of each cycle, we offer "optimal" bids at each location that are computed based on the current estimates of the underlying parameters at all the locations. In Broder and Rusmevichientong (2012), the corresponding notion is that of offering prices based on the current estimate of the underlying parameters of the demand curve; each price is the solution of a single-period revenue-maximization problem. In contrast, we face a capacity constraint across all locations – namely, that the number of impressions assigned to each campaign cannot exceed its requirement. Also, the number of additional impressions required for each campaign changes dynamically as impressions are won and allocated to the campaigns. Therefore, to compute the optimal bid in each time period at each location, we need to solve a stochastic dynamic program (DP) based on the current values of the estimated parameters at all the locations. Consequently, our analysis of the regret in the exploitation phase is necessarily more sophisticated than the one in Broder and Rusmevichientong (2012).

Broder and Rusmevichientong (2012) show that the regret under any pricing policy is $\Omega(\sqrt{N})$. If the total number of impressions required over all the campaigns in each season is no less than the number of periods in the season, then we also show an $\Omega(\sqrt{I})$ lower bound on the regret under any policy. If the total number of required impressions in each season is strictly less than the number of periods in the season, then we establish an $\Omega(I^{2/7})$ lower bound on the regret under any policy.

- den Boer and Zwart (2015) obtain an $\Omega(\log I)$ lower bound on the regret under any policy when the initial inventory in each season is strictly less than the number of periods in the season. They propose a pricing policy that is a modification of the certainty-equivalent pricing strategy (i.e., offering the optimal price with respect to the current parameter estimates); this policy achieves a regret of $\mathcal{O}(\log^2(I))$. However, we show that this bound cannot be achieved for our problem; in particular, as mentioned above, we establish an $\Omega(I^{2/7})$ lower bound on the regret under any policy when the required number of impressions in each season is strictly less than the number of periods in the season. The underlying reason for this difference is as follows. den Boer and Zwart (2015) show that their problem satisfies an "endogenous learning" property: if the chosen selling prices are sufficiently close to the optimal ones, then the unknown parameters can be learned fast. In turn, this is possible due to a minimum amount of price dispersion guaranteed in the optimal policy for the clairvoyant problem. However, in our context, no such bid dispersion is guaranteed since the arrival of impressions is uncertain. For instance, it is possible that no impression arrives during the early stages of a season and, as a result, when the first impression arrives, the maximum number of impressions that can be potentially acquired during the remainder of the season

is too few to exceed the targets of the ongoing campaigns in that season. In this case, the clairvoyant problem is effectively unconstrained and the optimal bids simply minimize the cost in each period and are therefore identical. Thus, since no bid dispersion can be guaranteed, our problem does not satisfy the endogenous learning property and the $\mathcal{O}(\log^2(I))$ upper bound on the regret cannot be achieved. A consequence of this is that our bidding policy requires active experimentation.

## Appendix L Technical Details in Section 5.2: Approximations of the DPs in the Learning Algorithm

Since the FEFS property (Section 2) no longer holds in the generalized setting where the unit penalty cost and the desired set of geographical locations (from where impressions are sought) may differ across campaigns, we first formulate a generalized clairvoyant problem that incorporates the allocation decisions and define another problem (problem $\mathrm{P}_0$ below) by replacing the random impression arrivals and random outcome of the winning of impressions in the clairvoyant problem with their expectations. The optimal objective value of problem $\mathrm{P}_0$ is a lower bound on that of the clairvoyant problem. Next, we present an alternate (and equivalent) formulation, defined as problem $\mathrm{P}_1$ below. The decisions of problem $\mathrm{P}_0$ are the bid prices at each location and the allocation decisions of the acquired impressions to the campaigns, in each time period. In the alternate formulation ($\mathrm{P}_1$), the decisions in each time period are the probabilities of winning impressions from each location, and the probabilities of winning impressions from each location and allocating them to specific campaigns. We then consider a problem – defined as problem $\mathrm{P}_2$ below – where the campaign-allocation decision for an acquired impression is randomized, and show that the optimal objective value of this problem is lower than that of problem $\mathrm{P}_1$. Next, we define a deterministic problem – problem ($\mathrm{P}_3$) below – by replacing the random campaign arrivals with their expectations, and show that the optimal objective value of problem $\mathrm{P}_3$ is a lower bound on that of $\mathrm{P}_2$. Finally, we consider another deterministic problem – defined as problem $\mathrm{P}_4$ below – that is related to problem $\mathrm{P}_3$ via the following property: Every feasible solution to $\mathrm{P}_3$ corresponds to a feasible solution to $\mathrm{P}_4$ with a (weakly) lower objective function value. Thus, the optimal objective value of problem $\mathrm{P}_4$ is a lower bound on that of problem $\mathrm{P}_3$. We now formally define problems $\mathrm{P}_0$–$\mathrm{P}_4$.

Consider the clairvoyant problem in season $i$. Let $\mathfrak{T}$ denote the number of weeks in each season and $\mathfrak{S}$ denote the number of time periods in each week. Then, the total number of time periods in each season is $T = \mathfrak{T} \cdot \mathfrak{S}$. In our numerical experiments, $\mathfrak{T} = 2$ and $\mathfrak{S} = 10^6$. For $i \in \{1, \cdots, I\}$, $\mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}$, and $\mathfrak{s} \in \{1, \cdots, \mathfrak{S}\}$, let $\zeta_{i,\mathfrak{t},\mathfrak{s}} = l$ if an impression arrives from location $l \in \mathcal{L}$ in period $(i, \mathfrak{t}, \mathfrak{s})$ and $\zeta_{i,\mathfrak{t},\mathfrak{s}} = 0$ if no impression arrives in that period. Thus, $\mathbb{E}[\mathbb{1}\{\zeta_{i,\mathfrak{t},\mathfrak{s}} = l\}] = q_l$, where $q_l$ is the arrival probability of an impression from location $l$ in each time period. Let $b^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} \in B$ denote the bid price at location $l$ in time period $(i, \mathfrak{t}, \mathfrak{s})$ under policy $\pi$. Let $d_{i,\mathfrak{t},\mathfrak{s}} = 1$ if the impression is won by bidding an amount $b^{\pi}_{i,\mathfrak{t},\mathfrak{s},\zeta_{i,\mathfrak{t},\mathfrak{s}}}$ at location $\zeta_{i,\mathfrak{t},\mathfrak{s}}$, and $d_{i,\mathfrak{t},\mathfrak{s}} = 0$ otherwise. Recall that $p_l(\gamma_l, b_l)$ is the *win curve* at location $l \in \mathcal{L}$, i.e., the probability of winning

an impression that arrives from location $l$ by bidding an amount $b_l \in B$, where $\gamma_l$ is a vector of parameters that characterize this distribution. Then, $d_{i,\mathfrak{t},\mathfrak{s}}$ is Bernoulli distributed with mean $p_{\zeta_{i,\mathfrak{t},\mathfrak{s}}}(\gamma_{\zeta_{i,\mathfrak{t},\mathfrak{s}}}, b^{\pi}_{i,\mathfrak{t},\mathfrak{s},\zeta_{i,\mathfrak{t},\mathfrak{s}}})$. Let $c$ denote a campaign with penalty cost $e_c$ for each unmet impression that starts at the beginning of week $\bar{\mathfrak{t}}_c$ and ends at the end of week $\underline{\mathfrak{t}}_c$, and requires impressions from locations in $\mathcal{L}_c$. Let $W_c$ be a random variable, with a known distribution, representing the number of impressions required by campaign $c$. We allow $\mathbb{P}(W_c = 0) > 0$, i.e., there is a positive probability that campaign $c$ does not arrive at the beginning of week $\bar{\mathfrak{t}}_c$. Let $\mathcal{C}$ denote the set of all possible campaigns. Let $(a^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} : l \in \mathcal{L})$ denote the allocation decision in time period $(i, \mathfrak{t}, \mathfrak{s})$ under policy $\pi$, with $a^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} = c \in \mathcal{C}$ indicating that if an impression arrives from location $l$ in that period and is won, then it is allocated to campaign $c$. Then, the optimization problem in season $i$ is:

$$\min_{\pi} \mathbb{E}\Bigg[ \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{\mathfrak{s}=1}^{\mathfrak{S}} \sum_{l \in \mathcal{L}} b^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} \mathbb{1}\{\zeta_{i,\mathfrak{t},\mathfrak{s}} = l\} d_{i,\mathfrak{t},\mathfrak{s}} + \sum_{c \in \mathcal{C}} \left( W_c - \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{\mathfrak{s}=1}^{\mathfrak{S}} \sum_{l \in \mathcal{L}} \mathbb{1}\{\zeta_{i,\mathfrak{t},\mathfrak{s}} = l\} d_{i,\mathfrak{t},\mathfrak{s}} \mathbb{1}\{a^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} = c\} \mathbb{1}\{\bar{\mathfrak{t}}_c \le \mathfrak{t} \le \underline{\mathfrak{t}}_c\} \mathbb{1}\{l \in \mathcal{L}_c\} \right)^{+} e_c \Bigg].$$

We replace the random impression arrivals (i.e., $\mathbb{1}\{\zeta_{i,\mathfrak{t},\mathfrak{s}} = l\}$) and the random outcome of the winning of impressions (i.e., $d_{i,\mathfrak{t},\mathfrak{s}}$) in the objective function above with their expectations to define problem $\mathrm{P}_0$ below. Note that the objective function of the clairvoyant problem above is convex in $\mathbb{1}\{\zeta_{i,\mathfrak{t},\mathfrak{s}} = l\}$ and $d_{i,\mathfrak{t},\mathfrak{s}}$. Thus, by Jensen's inequality, the optimal objective value of $\mathrm{P}_0$ is an lower bound on that of the clairvoyant problem.

$$\min_{\pi} \mathbb{E}\Bigg[ \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{\mathfrak{s}=1}^{\mathfrak{S}} \sum_{l \in \mathcal{L}} b^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} q_l p_l(\gamma_l, b^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}) + \sum_{c \in \mathcal{C}} \left( W_c - \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{\mathfrak{s}=1}^{\mathfrak{S}} \sum_{l \in \mathcal{L}} q_l p_l(\gamma_l, b^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}) \mathbb{1}\{a^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} = c\} \mathbb{1}\{\bar{\mathfrak{t}}_c \le \mathfrak{t} \le \underline{\mathfrak{t}}_c\} \mathbb{1}\{l \in \mathcal{L}_c\} \right)^{+} e_c \Bigg]. \tag{$\mathrm{P}_0$}$$

Next, we formulate an optimization problem that is equivalent to $\mathrm{P}_0$. In this problem, the two decisions in each time period are the probabilities of winning impressions that arrive from each location, and the probabilities of winning impressions from each location and allocating them to specific campaigns. Let $y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}$ denote the winning probability under policy $\pi$ at location $l$ in time period $(i, \mathfrak{t}, \mathfrak{s})$. Let $z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}$ denote the probability with which policy $\pi$ wins the impression from location $l$ and allocates it to campaign $c$ in time period $(i, \mathfrak{t}, \mathfrak{s})$. Note that an acquired impression can only be assigned to one of the campaigns. Thus, for $\mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}$, $\mathfrak{s} \in \{1, \cdots, \mathfrak{S}\}$, and $l \in \mathcal{L}$, only one of $(z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c} : c \in \mathcal{C})$ is positive, and that probability equals $y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}$. In the sequel, we drop $\gamma_l$ of $p_l(\gamma_l, b)$ when there is no ambiguity in doing so. Let $b_l(y)$ be the inverse of the function $p_l(b)$, i.e. $b_l(y) = p_l^{-1}(y)$. Thus, $b_l(y)$ is the bid price required to ensure a winning probability of $y$ at location $l$ in each time period. Let $\mathfrak{f}_l(y) := b_l(y)y$ denote the expected bidding cost associated with a target winning probability of $y$ for an impression at location $l$ in each time period. We show that $\mathfrak{f}_l(y)$ is convex in $y$ (see the proof of Lemma A.18). Let $y_l^{\min} = p_l(b^{\min})$ and $y_l^{\max} = p_l(b^{\max})$. Then, problem $\mathrm{P}_0$ can be equivalently written

as:

$$\mathfrak{f}^i := \min_{\pi} \mathbb{E}\left[\sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l \mathfrak{f}_l(y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}) + \sum_{c\in\mathcal{C}}\left(W_c - \sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}\right)^+ e_c\right]. \tag{P$_1$}$$

$$\text{s.t. } y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} = \sum_{c\in\mathcal{C}} z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L},$$

$$(y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} - z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}) z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c} = 0, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C},$$

$$z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c} \le \mathbb{1}\{\bar{\mathfrak{t}}_c \le \mathfrak{t} \le \underline{\mathfrak{t}}_c\}\mathbb{1}\{l\in\mathcal{L}_c\}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C},$$

$$y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} \in [y_l^{\min}, y_l^{\max}], \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}.$$

Consider now a relaxed version of problem $\mathrm{P}_1$ obtained by randomizing the impression-allocation decision, i.e., by allowing an acquired impression from one location to be assigned to multiple campaigns with positive probabilities. This is problem ($\mathrm{P}_2$) defined below. Clearly, the optimal objective value of problem ($\mathrm{P}_2$) is no greater than that of problem ($\mathrm{P}_1$).

$$\tilde{\mathfrak{f}}^i := \min_{\pi} \mathbb{E}\left[\sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l \mathfrak{f}_l(y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}) + \sum_{c\in\mathcal{C}}\left(W_c - \sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}\right)^+ e_c\right]. \tag{P$_2$}$$

$$\text{s.t. } y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l} = \sum_{c\in\mathcal{C}} z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L},$$

$$z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c} \le \mathbb{1}\{\bar{\mathfrak{t}}_c \le \mathfrak{t} \le \underline{\mathfrak{t}}_c\}\mathbb{1}\{l\in\mathcal{L}_c\}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C},$$

$$y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}, z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c} \in [y_l^{\min}, y_l^{\max}], \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C}.$$

Next, we construct a deterministic problem whose optimal objective value is an lower bound on that of problem ($\mathrm{P}_2$) (see proof of Lemma A.18). In this problem, the decisions in each time period are deterministic; thus, for convenience, we drop the policy superscript $\pi$ from $y^{\pi}_{i,\mathfrak{t},\mathfrak{s},l}$ and $z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}$. Let $\check{y} = (y_{i,\mathfrak{t},\mathfrak{s},l} : \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L})$ and $\check{z} = (z_{i,\mathfrak{t},\mathfrak{s},l,c} : \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, s\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C})$. Then, the following deterministic problem is obtained by replacing $W_c$ with its expectation:

$$\mathfrak{h}^i := \min_{\check{y},\check{z}} \left\{\sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l \mathfrak{f}_l(y_{i,\mathfrak{t},\mathfrak{s},l}) + \sum_{c\in\mathcal{C}}\left(\mathbb{E}[W_c] - \sum_{\mathfrak{t}=1}^{\mathfrak{T}}\sum_{\mathfrak{s}=1}^{\mathfrak{S}}\sum_{l\in\mathcal{L}} q_l z_{i,\mathfrak{t},\mathfrak{s},l,c}\right)^+ e_c\right\}. \tag{P$_3$}$$

$$\text{s.t. } y_{i,\mathfrak{t},\mathfrak{s},l} = \sum_{c\in\mathcal{C}} z_{i,\mathfrak{t},\mathfrak{s},l,c}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}$$

$$z_{i,\mathfrak{t},\mathfrak{s},l,c} \le \mathbb{1}\{\bar{\mathfrak{t}}_c \le \mathfrak{t} \le \underline{\mathfrak{t}}_c\}\mathbb{1}\{l\in\mathcal{L}_c\}, \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C}$$

$$y_{i,\mathfrak{t},\mathfrak{s},l}, z_{i,\mathfrak{t},\mathfrak{s},l,c} \in [y_l^{\min}, y_l^{\max}], \forall \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, \mathfrak{s}\in\{1,\cdots,\mathfrak{S}\}, l\in\mathcal{L}, c\in\mathcal{C}$$

Finally, we define another deterministic problem whose optimum objective value is a lower bound on that of problem $\mathrm{P}_3$. Let $y_{i,\mathfrak{t},l}$ denote the winning probability at location $l$ in each time period of week $\mathfrak{t}$ of season $i$. Let $z_{i,\mathfrak{t},l,c}$ denote the probability of winning the impression from location $l$ and allocating it to campaign $c$ in each time period of week $\mathfrak{t}$ of season $i$. Let $\hat{y} = (y_{i,\mathfrak{t},l} : \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, l\in\mathcal{L})$ and $\hat{z} = (z_{i,\mathfrak{t},l,c} : \mathfrak{t}\in\{1,\cdots,\mathfrak{T}\}, l\in$

$\mathcal{L}, c \in \mathcal{C}$). Consider the following deterministic problem:

$$\mathfrak{g}^i := \min_{(\hat{y},\hat{z})} \left\{ \mathfrak{S} \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{l \in \mathcal{L}} q_l \mathfrak{f}_l(y_{i,\mathfrak{t},l}) + \sum_{c \in \mathcal{C}} \left( \mathbb{E}[W_c] - \mathfrak{S} \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{l \in \mathcal{L}} q_l z_{i,\mathfrak{t},l,c} \right)^+ e_c \right\}. \tag{P$_4$}$$

$$\text{s.t. } y_{i,\mathfrak{t},l} = \sum_{c \in \mathcal{C}} z_{i,\mathfrak{t},l,c}, \forall \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, l \in \mathcal{L}$$

$$z_{i,\mathfrak{t},l,c} \leq \mathbb{1}\{\underline{\mathfrak{t}}_c \leq \mathfrak{t} \leq \bar{\mathfrak{t}}_c\} \mathbb{1}\{l \in \mathcal{L}_c\}, \forall \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C}$$

$$y_{i,\mathfrak{t},l}, z_{i,\mathfrak{t},l,c} \in [y_l^{\min}, y_l^{\max}], \forall \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C}$$

In the proof of the following result, we will show that the optimal objective value of problem (P$_4$) is a lower bound on that of problem (P$_3$). Thus, we have

**Lemma A.18** *The optimal objective value of problem* (P$_4$) *is a lower bound on that of problem* (P$_1$)*, i.e.,* $\mathfrak{g}^i \leq \mathfrak{f}^i$.

**Proof of Lemma A.18**: The conclusion that the optimal objective value of problem (P$_2$) is a lower bound on that of problem (P$_1$), i.e., $\mathfrak{f}^i \geq \tilde{\mathfrak{f}}^i$, is trivial. Further, since $\left(W_c - \sum_{\mathfrak{t}=1}^{\mathfrak{T}} \sum_{\mathfrak{s}=1}^{\mathfrak{S}} \sum_{l \in \mathcal{L}} q_l z^{\pi}_{i,\mathfrak{t},\mathfrak{s},l,c}\right)^+ e_c$ is convex in $W_c$, we have (from Jensen's inequality) that the optimal objective value of problem (P$_3$) is a lower bound on that of problem (P$_2$), i.e., $\tilde{\mathfrak{f}}^i \geq \mathfrak{h}^i$.

Next, we show that the optimal objective value of problem (P$_4$) is a lower bound on that of problem (P$_3$), i.e., $\mathfrak{h}^i \geq \mathfrak{g}^i$. Let $(y^*_{i,\mathfrak{t},\mathfrak{s},l}, z^*_{i,\mathfrak{t},\mathfrak{s},l,c} : \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, s \in \{1, \cdots, \mathfrak{S}\}, l \in \mathcal{L}, c \in \mathcal{C})$ denote an optimal solution of problem (P$_3$). Let $\hat{y}_{i,\mathfrak{t},l} = \frac{\sum_{\mathfrak{s}=1}^{\mathfrak{S}} y^*_{i,\mathfrak{t},\mathfrak{s},l}}{\mathfrak{S}}$ and $\hat{z}_{i,\mathfrak{t},l,c} = \frac{\sum_{\mathfrak{s}=1}^{\mathfrak{S}} z^*_{i,\mathfrak{t},\mathfrak{s},l,c}}{\mathfrak{S}}$. It is straightforward to verify that $(\hat{y}_{i,\mathfrak{t},l}, \hat{z}_{i,\mathfrak{t},l,c} : \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C})$ is a feasible solution to problem (P$_4$). In addition, $\mathfrak{f}_l(y)$ is convex in $y$: Recall that $p_l(b)$ is log concave in $b$, which implies that $\frac{\partial \log p_l(b)}{\partial b} = \frac{\partial p_l(b)/\partial b}{p_l(b)}$ decreases in $b$. Then,

$$\frac{\partial \mathfrak{f}_l(y)}{\partial y} = b_l(y) + y \frac{\partial b_l(y)}{\partial y}$$

increases in $y$. It is clear that $b_l(y)$ increases in $y$. Note that $y \frac{\partial b_l(y)}{\partial y} = \frac{p_l(b)}{\partial p_l(b)/\partial b}$ increases in $y$. Thus, $\mathfrak{f}_l(y)$ is convex in $y$ and the objective function of problem (P$_3$) is convex in $(\check{y}, \check{z})$. Consequently, the objective value of problem (P$_4$) under $(\hat{y}_{i,\mathfrak{t},l}, \hat{z}_{i,\mathfrak{t},l,c} : \mathfrak{t} \in \{1, \cdots, \mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C})$ is a lower bound of the optimal objective value of problem (P$_3$). ■

Lemma A.18 immediately implies that the regret of any policy $\pi$ is smaller than the difference between the expected cost under policy $\pi$ and $\sum_{i=1}^{I} \mathfrak{g}^i$.

Recall that at the beginning of each exploitation phase, to compute the optimal bid based on the latest estimates, we need to solve a DP with a multi-dimensional state space. Similar to the convex optimization problem we developed above to approximate the DP of the clairvoyant problem above, we solve a convex optimization problem in our numerical experiments to approximate the DP in the exploitation phase. Consider an exploitation phase that starts in time period $(i, \mathfrak{t}, \mathfrak{s})$. Let $\check{\mathcal{C}}_{i,\mathfrak{t},\mathfrak{s}}$ denote the set of campaigns that arrive

in or before time period $(i,\mathfrak{t},\mathfrak{s})$. Let $\hat{\mathcal{C}}_{i,\mathfrak{t},\mathfrak{s}}$ denote the set of campaigns that arrive after time period $(i,\mathfrak{t},\mathfrak{s})$. For any $c \in \check{\mathcal{C}}_{i,\mathfrak{t},\mathfrak{s}}$, let $w_c$ denote the total number of unmet impressions for campaign $c$ in period $(i,\mathfrak{t},\mathfrak{s})$. Let $\hat{\mathfrak{f}}_l(y)$ denote the expected bidding cost associated with a target win probability of $y$ for an impression at location $l$ in each time period based on the updated estimates. Let $\tilde{y} = (y_{i,\hat{\mathfrak{t}},l} : \hat{\mathfrak{t}} \in \{\mathfrak{t},\cdots,\mathfrak{T}\}, l \in \mathcal{L})$ and $\tilde{z} = (z_{i,\hat{\mathfrak{t}},l,c} : \hat{\mathfrak{t}} \in \{\mathfrak{t},\cdots,\mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C})$. Then, we solve the following problem to approximate the DP.

$$\min_{\tilde{y},\tilde{z}} \left\{ \begin{array}{l} (\mathfrak{S}-\mathfrak{s}+1)\sum_{l\in\mathcal{L}} q_l \hat{\mathfrak{f}}_l(y_{i,\mathfrak{t},l}) + \mathfrak{S}\sum_{\hat{\mathfrak{t}}=t+1}^{\mathfrak{T}}\sum_{l\in\mathcal{L}} q_l \hat{\mathfrak{f}}_l(y_{i,\hat{\mathfrak{t}},l}) + \\ \sum_{c\in\check{\mathcal{C}}_{i,\mathfrak{t},\mathfrak{s}}} \left[ w_c - (\mathfrak{S}-\mathfrak{s}+1) q_l z_{i,\mathfrak{t},l,c} - \mathfrak{S}\sum_{\hat{\mathfrak{t}}=t+1}^{\mathfrak{T}}\sum_{l\in\mathcal{L}} q_l z_{i,\hat{\mathfrak{t}},l,c} \right]^+ e_c + \\ \sum_{c\in\hat{\mathcal{C}}_{i,\mathfrak{t},\mathfrak{s}}} \left[ \mathbb{E}[W_c] - (\mathfrak{S}-s+1) q_l z_{i,\mathfrak{t},l,c} - \mathfrak{S}\sum_{\hat{\mathfrak{t}}=t+1}^{\mathfrak{T}}\sum_{l\in\mathcal{L}} q_l z_{i,\hat{\mathfrak{t}},l,c} \right]^+ e_c \end{array} \right\} \tag{A-31}$$

$$\text{s.t. } y_{i,\hat{\mathfrak{t}},l} = \sum_{c\in\mathcal{C}} z_{i,\hat{\mathfrak{t}},l,c}, \forall \hat{\mathfrak{t}} \in \{\mathfrak{t},\cdots,\mathfrak{T}\}, l \in \mathcal{L}$$

$$z_{i,\hat{\mathfrak{t}},l,c} \le \mathbb{1}\{\bar{\mathfrak{t}}_c \le \hat{\mathfrak{t}} \le \underline{\mathfrak{t}}_c\}\mathbb{1}\{l \in \mathcal{L}_c\}, \forall \hat{\mathfrak{t}} \in \{\mathfrak{t},\cdots,\mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C}$$

$$y_{i,\hat{\mathfrak{t}},l}, z_{i,\hat{\mathfrak{t}},l,c} \in [y_l^{\min}, y_l^{\max}], \forall \hat{\mathfrak{t}} \in \{\mathfrak{t},\cdots,\mathfrak{T}\}, l \in \mathcal{L}, c \in \mathcal{C}$$

The optimal solution of the above problem is used to obtain the bid price and allocation decision during the exploitation phase under our policy. The expected cost under these bid prices and allocation decisions is greater than the optimal cost-to-go of the DP based on the updated estimates. Thus, the difference between the expected cost under our policy, where the bid prices and allocation decisions during the exploitation phase are obtained by solving problem (A-31) above, and $\sum_{i=1}^{I} \mathfrak{g}^i$ is an upper bound on the true regret under our policy.

## Appendix M Decomposition of Regret (Section 5.4): Details of Numerical Analysis

In this section, we discuss the details of our numerical study to address the two questions, defined in Section 5.4, on the decomposition of regret under our policy.

In order to compute the optimal cost of the clairvoyant problem, which is used as a benchmark to compute the true regret under our policy, we need to solve a DP with multi-dimensional state space optimally. Thus, we conduct another numerical study on a tractable scale. In this new setting, each season consists of 200 time periods. There are 3 locations indexed by $l = 1,2,3$ and the win curve at location $l$ is $p_l(\gamma_l, b_l) = \frac{\exp(\gamma_{l,1}+\gamma_{l,2}b_l)}{1+\exp(\gamma_{l,1}+\gamma_{l,2}b_l)}; l = 1,2,3$, where the true values of the parameters in the win curves are $(\gamma_{1,1},\gamma_{1,2}) = (-2.281, 0.705), (\gamma_{2,1},\gamma_{2,2}) = (-2.192, 1.042)$, $(\gamma_{3,1},\gamma_{3,2}) = (-1.905, 0.876)$, and $\Gamma_l^{(0)} = [-8,-0.1]\times[0.1,8]$, $l = 1,2,3$. Impressions acquired from any of these locations can be used to satisfy the requirement of any campaign. The duration of a campaign is either 100 periods or 200 periods, and each campaign requires 10 impressions. The penalty cost of each unmet impression can take two values: 10 and 15. If the duration of a campaign is one week (resp., two weeks), then the unit penalty cost is 10 (resp., 15). In each time period, an impression arrives from location $l \in \{1,2,3\}$ with probability 0.1. Campaigns can arrive at the beginning of the first period or the $101^{st}$ period. Consider an arbitrary season: At most one campaign can arrive at the beginning of the first period (resp., $101^{st}$ period).

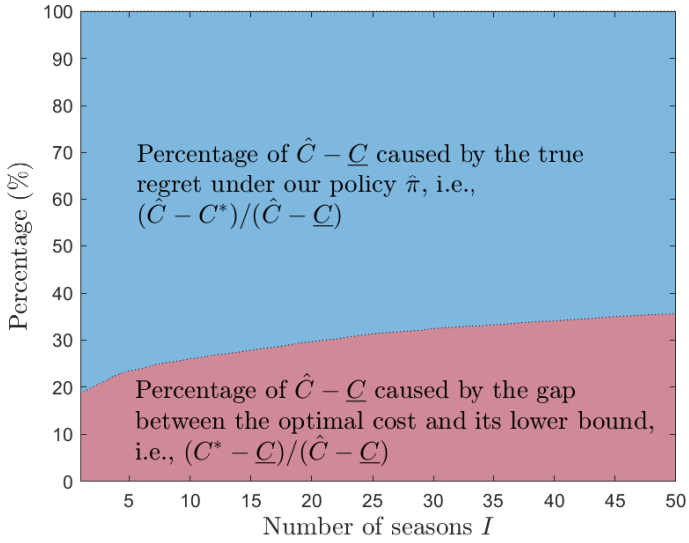
At the beginning of the first period, either a 100-period or a 200-period campaign can arrive; each of these two events occurs with probability 0.48. At the beginning of the $101^{st}$ period, a 100-period campaign arrives with probability 0.96.

Recall from Section 5.2 that, to approximate the DP of the clairvoyant problem, we solve a convex optimization problem whose optimal objective value is a lower bound on the optimal cost of the clairvoyant problem. Thus, our (reported) regret (i.e., the difference between the expected cost under our policy and the lower bound on the optimal cost of the clairvoyant problem) is, in fact, an upper bound on the "true regret" under our policy. Let $C^*$ and $\underline{C}$ denote the optimal cost of the clairvoyant problem and its lower bound, respectively. Further, for the general setting where the unit penalty costs differ across campaigns, when the underlying parameters of the win curves are unknown, the FEFS allocation policy is no longer optimal. We instead define and solve a convex optimization problem (based on the estimates of the parameters of the win curves) rather than solving the DP in the exploitation phase of our policy. The optimal solution of this problem is then used to obtain the bid price and allocation decision. Let $\hat{\pi}$ denote our policy and let $\hat{C}$ denote the expected cost under $\hat{\pi}$.

Note that $\hat{C} - \underline{C} = (\hat{C} - C^*) + (C^* - \underline{C})$. We first examine how much of the difference between $\hat{C}$ and $\underline{C}$ is the true regret under our policy, i.e., the difference between $\hat{C}$ and $C^*$, and how much of it is caused by the gap between the optimal cost $C^*$ and its lower bound $\underline{C}$. For each value of $I$, we compute the average cost under our policy over 100 simulations. Figure 6 plots the percentage of the difference between $\hat{C}$ and $\underline{C}$ that is caused by the "true" regret under our policy, i.e., $\hat{C} - C^*$, and the percentage caused by the gap between the optimal cost and its lower bound, i.e., $C^* - \underline{C}$, as a function of the number of seasons $I$. After the first season, about 81% of the difference between $\hat{C}$ and $\underline{C}$ is caused by the true regret under our policy and 19% is caused by the gap between $C^*$ and $\underline{C}$. As time goes by, the true regret under our policy per season reduces while the gap (per season) between the optimal cost and its lower bound remains unaffected. After 50 seasons, about 64% of the difference between $\hat{C}$ and $\underline{C}$ is caused by the true regret under our policy and 36% is caused by the gap between $C^*$ and $\underline{C}$. Therefore, a substantial portion of the loss under our policy compared to the lower bound of the optimal cost is due to the gap between the optimal cost of the clairvoyant problem and its lower bound.

Next, we examine how much of the (true) regret (i.e., the difference between the expected cost $\hat{C}$ under our policy $\hat{\pi}$ and the optimal cost $C^*$ of the clairvoyant problem) is caused by learning and how much of it is caused by the approximation of the DP (note that this includes the regret caused due to the possible suboptimal allocation of impressions when the parameters of the win-curve are known). Consider a policy $\tilde{\pi}$ in which the bid price and allocation decision in each period are obtained by solving the convex optimization problem optimally for the clairvoyant problem (i.e., one in which the platform has full information about the win curves at all the locations in advance). Let $\tilde{C}$ denote the expected cost under policy $\tilde{\pi}$. Then, the regret under our policy $\hat{\pi}$ equals $\hat{C} - C^* = (\hat{C} - \tilde{C}) + (\tilde{C} - C^*)$. The first term, namely $\hat{C} - \tilde{C}$, is the regret caused by learning, while the second term, namely $\tilde{C} - C^*$, is the regret caused by the approximation of the DP (including the suboptimal

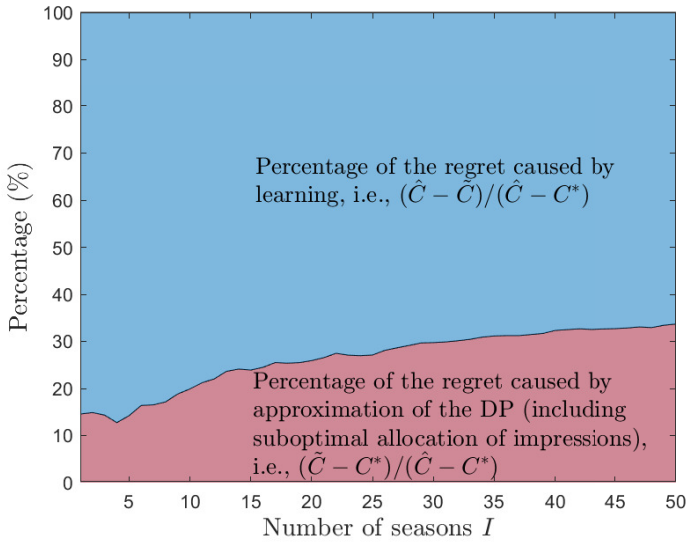
Figure 6: Percentage of the difference between $\hat{C}$ and $\underline{C}$ caused by the true regret under our policy and the gap between the optimal cost and its lower bound




$\hat{C}$: the average cost under our policy $\hat{\pi}$; $C^*$: the optimal cost of the clairvoyant problem; $\underline{C}$: the lower bound on the optimal cost of the clairvoyant problem.

allocation of impressions). For each value of $I$, we compute the average cost under our policy $\hat{\pi}$ and the average cost under policy $\tilde{\pi}$ over 100 simulations. Figure 7 plots the percentage of the regret caused by learning (resp., suboptimal allocation of impressions) as a function of the number of seasons ($I$). After the first season, about 85% of the regret is caused by learning and at most 15% of the regret is caused by the suboptimal allocation of impressions. With time, the learning of the win curves improves and we get progressively better estimates of the parameters of the win curves. Therefore, the percentage of the regret caused by learning reduces over time. After 50 seasons, about 66% of the regret is caused by learning and at most 34% is caused by the suboptimal allocation of impressions. To summarize, learning (in other words, insufficient knowledge of the parameters of the win curves) is the dominant cause of regret.

Figure 7: Percentage of the regret caused by learning and approximation of the DP (including suboptimal allocation of impressions).




$\hat{C}$: the average cost under our policy $\hat{\pi}$; $C^*$: the optimal cost of the clairvoyant problem; $\tilde{C}$: the average cost under policy $\tilde{\pi}$.