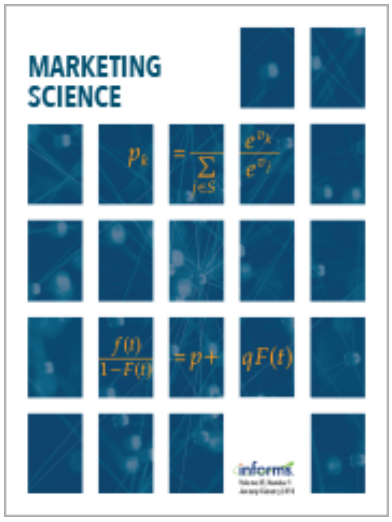


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimizing User Engagement Through Adaptive Ad Sequencing

Omid Rafieian

To cite this article:
Omid Rafieian (2023) Optimizing User Engagement Through Adaptive Ad Sequencing. Marketing Science 42(5):910-933. https://doi.org/10.1287/mksc.2022.1423

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 42, No. 5, September–October 2023, pp. 910–933
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Optimizing User Engagement Through Adaptive Ad Sequencing

Omid Rafieian[a,b]

[a] Cornell Tech, New York, New York 10044; [b] SC Johnson College of Business, Cornell University, Ithaca, New York 14853
Contact: or83@cornell.edu, https://orcid.org/0000-0001-8633-2302 (OR)

Received: February 3, 2022
Revised: August 15, 2022
Accepted: October 21, 2022
Published Online in Articles in Advance: December 29, 2022

https://doi.org/10.1287/mksc.2022.1423

Abstract. In this paper, we propose a unified dynamic framework for adaptive ad sequencing that optimizes user engagement with ads. Our framework comprises three components: (1) a Markov decision process that incorporates intertemporal tradeoffs in ad interventions, (2) an empirical framework that combines machine learning methods with insights from causal inference to achieve personalization, counterfactual validity, and scalability, and (3) a robust policy evaluation method. We apply our framework to large-scale data from the leading in-app ad network of an Asian country. We find that the dynamic policy generated by our framework improves the current practice in the industry by 5.76%. This improvement almost entirely comes from the increased average ad response to each impression instead of the increased usage by each user. We further document a U-shaped pattern in improvements across the length of the user's history, with high values when the user is new or when enough data are available for the user. Next, we show that ad diversity is higher under our policy and explore the reason behind it. We conclude by discussing the implications and broad applicability of our framework to settings where a platform wants to sequence content to optimize user engagement.

History: Olivier Toubia served as the senior editor for this article.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mksc.2022.1423.

Keywords: advertising • personalization • adaptive interventions • policy evaluation • dynamic programming • machine learning • offline reinforcement learning

1. Introduction

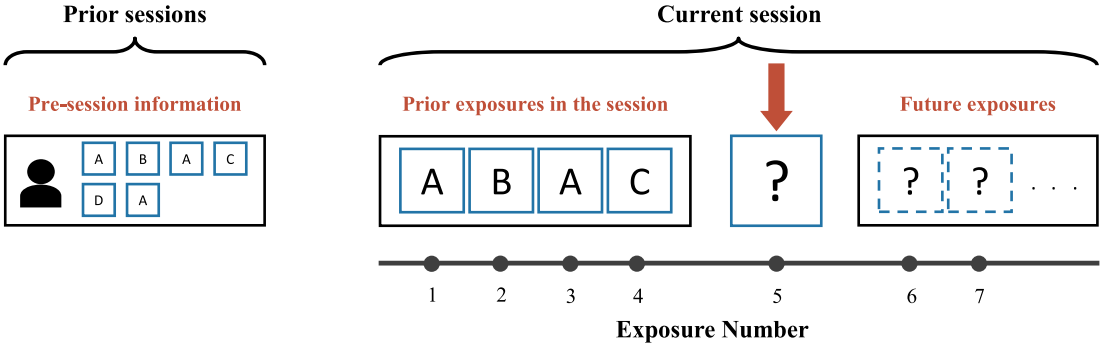
1.1. Motivation for Adaptive Ad Sequencing

Consumers now spend a significant portion of their time on mobile apps. The average time spent on mobile apps by U.S. adults has grown steadily over the last few years, surpassing four hours per day for the first time in the first quarter of 2021 (Kristianto 2021). This demand expansion, in turn, has amplified marketing activities targeted toward mobile app users. In 2020, mobile advertising generated nearly $100 billion in the United States, accounting for more than double the share of its digital counterpart, desktop advertising (IAB 2021). Most of this growth in mobile advertising is attributed to in-app ads (i.e., ads shown inside mobile apps), with more than 80% of ad spend in the mobile advertising category (eMarketer 2018).

Two key features of mobile in-app ads have contributed to this dramatic growth. First, the mobile app ecosystem has excellent user tracking ability, thereby allowing "personalization" of ad interventions and targeting of users based on their prior behavioral history (Han et al. 2012). Second, in-app ads are usually refreshable and dynamic in nature: Each ad intervention is shown for a fixed amount of time (e.g., 30 seconds or one minute) inside the app and followed by another ad intervention. As such, a user can see multiple ad exposures within a session.[1] Refreshable ads, together with the potential for personalization, make in-app advertising amenable to "adaptive ad sequencing," that is, optimizing the sequence of ads based on real-time behavioral information.

Adaptive ad sequencing brings a forward-looking perspective to the publisher's ad allocation problem.[2] That is, sequencing not only captures the immediate user engagement when making a decision to show an ad based on the information available, but it also takes user engagement in future events and exposures into account. Figure 1 illustrates this point by differentiating between the information available from the past and the information that would be available in the future. However, most platforms do not use a forward-looking model for ad allocation because it substantially adds to the complexity of the model.[3] This is one of the reasons why the current state of advertising practice is to use supervised learning and contextual bandit algorithms that only focus on the data available at the moment and ignore the future exposures (Theocharous et al. 2015). Furthermore, the returns from adopting a forward-looking model are not clear. Thus, the publisher's decision on whether to use a dynamic framework boils

Figure 1. (Color online) Visual Schema for the Publisher's Ad Sequencing Decision



Notes. The user is at the fifth exposure in the session, and the publisher needs to decide which ad to show to this user. Unlike the myopic publisher that only uses the information from the past, a forward-looking publisher also accounts for the futures exposures when making the decision.

down to whether incorporating future information helps them achieve a better outcome.

In principle, using a forward-looking framework is only valuable when there is interdependence between ad exposures; that is, the ad shown in the current exposure affects the performance of future exposures and the overall value created by the system. The impact of the publisher's current decision on the future exposures can fundamentally be of two types: (1) extensive margin, which means that the user will stay longer in the session and generate more exposures, and (2) intensive margin, which means that the engagement with each exposure in the future will be higher, on average. Prior literature on advertising offers multiple accounts that suggest a great possibility for value creation through both channels. On the one hand, sequencing can result in greater usage in light of studies on the link between advertising and subsequent usage (Wilbur 2008, Goli et al. 2021). On the other hand, sequencing can increase the response rate to each exposure by better managing effects of carryover, spillover, temporal spacing, and variety (Rutz and Bucklin 2011, Jeziorski and Segal 2015, Sahni 2015, Lu and Yang 2017, Rafieian and Yoganarasimhan 2022b).

1.2. Research Agenda and Challenges

The dynamic effects of advertising give rise to the intertemporal tradeoffs in ad allocation. For example, Rafieian and Yoganarasimhan (2022b) find that an increase in the variety of ads in a session results in a higher engagement with the next ad. However, it is not clear that increasing variety is the optimal decision at any point because it can come at the expense of showing an irrelevant ad. Although the dynamic effects of advertising and the resulting intertemporal tradeoffs are well established in the literature, neither research nor practice has looked into how we can collectively incorporate these findings to optimize publisher's outcomes by dynamically sequencing ads. Our goal in this paper is to fill this gap by developing a unified framework for adaptive ad sequencing and documenting the gains from this framework.

To build such a framework, we first need to specify our objective. We view the problem through the lens of a publisher who aims to maximize the expected number of clicks per session. Although our framework is general and can accommodate any measure of user engagement over any optimization horizon, we focus on clicks as our measure of user engagement because clicks are instrumental to a publisher's business model in mobile in-app advertising. With our objective in place, we seek to answer the following three questions:

1. How can we develop a unified dynamic framework that incorporates the intertemporal tradeoffs in ad allocation and designs a policy that maximizes user engagement?

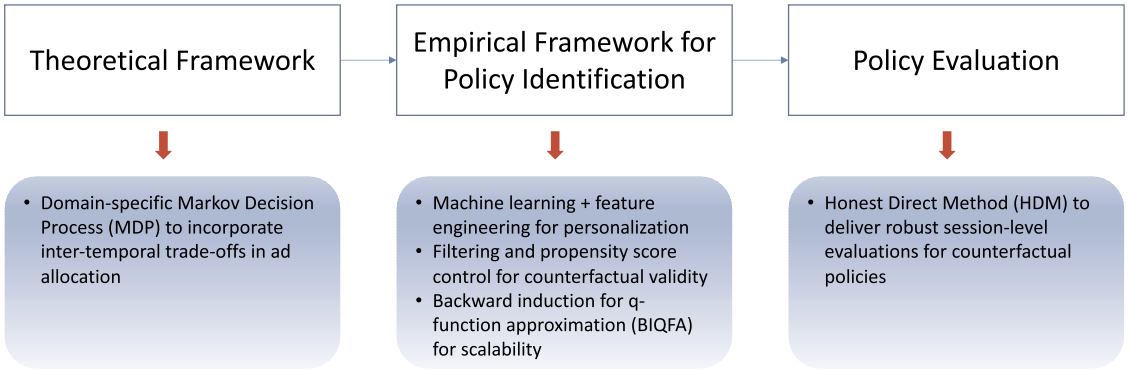
2. How can we empirically evaluate the performance of the counterfactual policy identified by our adaptive ad sequencing framework?

3. What are the gains from using our adaptive ad sequencing framework over existing benchmarks? Are these gains due to increased usage (extensive margin) or increased average ad response (intensive margin)? Which session characteristics are linked to greater gains? How different is the policy identified by our framework from the benchmark policies?

1.3. Our Approach

In this paper, we present a unified three-pronged framework that addresses these challenges and develops an adaptive ad sequencing policy to maximize user engagement with ads. We present an overview of our approach in Figure 2, where the top row illustrates that we start with a theoretical framework that models the domain structure of our problem and informs us of the key empirical tasks required for policy identification and evaluation, and the bottom row describes the specifics of our approach.

Figure 2. (Color online) Overview of Our Approach



For our theoretical framework, we specify a domain-specific Markov decision process (MDP) that characterizes the structure of adaptive ad interventions. In particular, we use a rich set of state variables that collectively incorporate the dynamic effects of advertising identified in the literature. Our MDP characterizes the reward at any exposure and how the state evolves in future periods, given any action taken by the publisher. Because our goal is to optimize the number of clicks per session, we define the reward as the expected probability of click, given the state variable and ad. This probability is also part of the state transition, as it helps us update the user's preference in real time for the next period. Another probabilistic factor that affects the future state is the expected probability of the user leaving the session after an intervention, which determines with what probability the user will be available to see the next ad exposure. The combination of reward and transition functions allows us to characterize the publisher's optimization problem theoretically.

Next, to empirically identify the optimal sequencing policy, we develop an empirical framework that allows us to evaluate all possible sequencing policies for each session. As broken down by our theoretical framework, we first need to obtain personalized estimates of the primitives of our MDP: expected click and leave probabilities. We do so by using machine learning methods that can capture more complex relationships between the covariates and the outcome. In particular, we use an extreme gradient boosting (XGBoost) algorithm with a rich set of features to predict click and leave outcomes. To ensure the counterfactual validity of our estimates, we use key insights from the causal inference literature and narrow down our focus on counterfactual sequences that *could have been shown* in our data. This is because machine learning algorithms can only generate accurate predictions for instances within the joint distribution of the training set used for model fitting. Furthermore, we control for propensity scores to account for potential selection in our predictions. Last, for the scalability of our empirical framework, we develop an algorithm called *backward induction for q-function approximation* (BIQFA) that takes the primitive estimates and learns a function that approximates the expected sum of current and future rewards for each pair of state variables and ad. This function approximation approach avoids the exhaustive search over any pair of state and ad, thereby reducing the computational burden substantially.

Although our empirical framework for policy identification separately evaluates each policy to find the optimal one, we cannot use the same evaluation approach to assess the performance of our policy because the policy identified by our framework will always outperform other policies by construction. To address this challenge, we develop an approach called *Honest Direct Method* (HDM) that completely separates the evaluation criteria for policy identification and policy evaluation: The data and models used for identification have no overlap with the data and model used for evaluation. To increase the robustness of this approach, we use a fully held-out third data set for final evaluation that is not used for model building in either policy identification or policy evaluation stages.

1.4. Findings and Contributions

We apply our framework to the data from a leading mobile in-app ad network of a large Asian country. Our setting has notable features that make it amenable to our research goals. First, the ad network uses a refreshable ad format where ad interventions last for one minute and change within the session. Second, the ad network runs a quasi-proportional auction that uses a probabilistic allocation rule, which induces a high degree of randomization in ad allocation Together, these two features create exogenous variation in the sequences shown in the data, thereby satisfying an essential requirement for our framework.

To establish the performance of our adaptive ad sequencing framework, we evaluate the gains from *adaptive ad sequencing policy* relative to three benchmark policies: (1) *random policy*, which selects ads randomly and is often used as a benchmark in the reinforcement

learning (RL) literature, (2) *single-ad policy*, which only shows a single ad with the highest reward in the session, thereby mimicking the practice of using a nonrefreshable ad slot as is common in desktop advertising, and (3) *adaptive myopic policy*, which uses all the information available and selects the ad with the highest reward at any exposure but ignores the expected future rewards. The adaptive myopic policy reflects the standard practice in the advertising industry, where publishers use supervised learning or contextual bandit algorithms to estimate click-through rate (CTR) for an ad in a given impression.

We evaluate all these policies on a completely held-out test set using different metrics. First, we document a 79.59% increase in the expected number of clicks from our fully dynamic policy over the random policy. Next, we show that our fully dynamic policy results in 27.46% greater expected number of clicks per session than the single-ad policy. This finding demonstrates the opportunity cost of using a nonrefreshable ad slot throughout the session, supporting the current industry trend of using refreshable ad slots. Finally, we focus on our key comparison in this paper and demonstrate a 5.76% gain in the expected number of clicks per session from our fully dynamic policy over the adaptive myopic policy. This suggests that choosing the best match at any point will not necessarily create the best match outcome at the end of the session. Instead, the right action sometimes is to show the ad that is not necessarily the best match at the moment but transitions the session to a better state in the future. This finding provides a strong proof-of-concept for the use of our framework. It has important implications for publishers and ad networks, especially because the current practice in the industry overlooks the dynamics of ad sequencing.

We further compare our policy with the benchmark policies using two other metrics: session length and ad concentration. Focusing on the session length shows how much of the gains from our policy come from an increase in usage and the number of impressions generated (extensive margin). Although our policy achieves a slightly higher session length, it is only 0.2% greater than the session length under the adaptive myopic policy, which suggests that the source for our gains is not the increase in usage but the increase in the average ad response rate (intensive margin). We then focus on ad concentration as our next metric and use the Herfindahl-Hirschman index (HHI) for ads shown under each policy. Our results reveal an interesting pattern: Adaptive ad sequencing policy results in a lower HHI than both adaptive myopic and single-ad policies, suggesting a greater ad diversity under our policy. A greater ad diversity can have long-term implications for the competition between advertisers and welfare impacts for consumers.

Next, to better interpret the mechanism underlying our gains, we explore the heterogeneity in gains from our policy over the adaptive myopic policy. We document a U-shaped pattern in gains over the number of prior sessions a user has been part of. This pattern suggests a mix of accounts as the user becomes more experienced that affect the gains in opposite directions. We explore these potential accounts in a series of regression models that illustrate the heterogeneity in gains across presession covariates (e.g., number of prior impressions or clicks by the user). To understand where the difference between our policy and adaptive myopic policy comes from, we first measure the discrepancy between the distribution of ad allocation under the two policies using different measures such as ℓ-norm and Kullback-Leibler divergence and then regress this discrepancy measure on presession characteristics. We find that a higher number of past impressions is associated with a greater discrepancy in distributions. In contrast, a higher variety of prior ads and number of past clicks are associated with a lower discrepancy in distributions.

In sum, our paper makes several contributions to the literature. First, from a methodological standpoint, we develop a unified dynamic framework that takes the past advertising data and scalably produces an optimal dynamic policy to personalize the sequence of ads in a session. A key contribution of our adaptive ad sequencing framework that comes from the use of the backward induction q-function approximation (BIQFA) algorithm is that it does not impose restrictive assumptions on the dynamic structure of the problem and remains agnostic about how dynamics arise in our setting. To our knowledge, this is the first paper that takes a prescriptive approach to generate an optimal dynamic policy by collectively incorporating the dynamic effects of advertising documented in the literature. Substantively, we establish the gains from our dynamic framework over a set of benchmarks that are often used in research and practice. In particular, we demonstrate that the gains from adopting the dynamic policy generated by our framework are 5.76%, compared with the adaptive myopic policy. This proof-of-concept is particularly important as the current practice in this industry uses the adaptive myopic policy and ignores the dynamics of the ad allocation problem. We further present a comprehensive analysis of the gains from our framework to provide interpretation for the mechanism underlying the gains. Our findings shed light on when and why our framework is more valuable than alternative policies. Last, from a managerial perspective, our framework is fairly general and can be a applied to a wide variety of domains where a platform or publisher aims to optimally sequence content to achieve better user-level outcomes, such as sequencing of articles to increase audience engagement with the content in news websites, sequencing of social media posts to increase user interaction and engagement, and sequencing of push notifications to reduce customer churn.

2. Related Literature

First, our paper relates to the marketing literature on personalization and targeting. Early papers in this stream build Bayesian frameworks that exploit behavioral data and personalize marketing mix variables (Rossi et al. 1996, Ansari and Mela 2003, Manchanda et al. 2006). Recent papers in this domain use machine learning algorithms often combined with insights from causal inference to achieve greater personalization in different domains such as search (Yoganarasimhan 2020), advertising (Rafieian and Yoganarasimhan 2021), free trial length (Yoganarasimhan et al. 2022), and product versioning through offering different ad loads to users (Goli et al. 2021).[4] Although all these papers focus on prescriptive or substantive frameworks to study personalization, they all study this phenomenon from a static point of view. Our paper extends this literature by bringing a dynamic objective to this problem and offering a scalable framework to develop forward-looking personalized targeting policies. From a substantive viewpoint, we show that the gains from adopting such forward-looking personalized policies is 5.76% compared with the baseline of myopic personalized policies.

Second, our work relates to both the substantive and prescriptive literature on the dynamics of advertising. Early work in this domain focuses on aggregate advertising models to understand ad responses over time and strategies such as pulsing (Horsky 1977, Little 1979, Simon 1982, Naik et al. 1998, Dubé et al. 2005, Aravindakshan and Naik 2011).[5] More recent papers in this domain use larger scale individual-level data of digital advertising and document different dynamic effects of advertising, such as effects of ad carryover or spillover, temporal spacing, and variety in search advertising (Rutz and Bucklin 2011, Jeziorski and Segal 2015, Sahni 2015, Lu and Yang 2017, Zantedeschi et al. 2017, Rafieian and Yoganarasimhan 2022b). Inspired by the dynamics of advertising, a different stream of work brings a more prescriptive view to the problem and focuses on the optimal policy design for advertisers and platforms. Given the complexity of the problem, these papers often simplify the problem by mapping the entire space into a few segments (Urban et al. 2013), ignoring intertemporal tradeoffs through a bandit specification (Schwartz et al. 2017), or imposing some structure on the dynamics to find a closed-form solution (Wilbur et al. 2013, Kar et al. 2015, Sun et al. 2017). Table 1 summarizes the prior work on ad allocation in terms of using (1) individual-level data to allow for ad personalization, (2) forward-looking (as opposed to myopic) framework, (3) high-dimensional state space that captures all the dynamic effects of advertising, and (4) no parametric assumption on state transitions (dynamics-agnostic). As shown in Table 1, none of the existing work of ad allocation satisfies all the four criteria, which highlights the contribution of our paper: Using the BIQFA algorithm allows us to collectively incorporate all the documented dynamic effects of advertising and find the optimal dynamic policy without reducing the richness and dimensionality of the state space or imposing any structure on the dynamics of the problem.

Finally, our paper relates to the literature on offline or batch RL, where the learner does not actively interact with the environment and must rely on observational data from the past to design an optimal dynamic policy. This class of problems is particularly relevant when safety guarantees are of utmost priority, and the system is not allowed to actively explore (Thomas et al. 2019). An important task in all these problems is to find a robust approach to evaluate counterfactual policies, that is, policies that have not necessarily been implemented in the data available. This problem is often referred to as off-policy policy evaluation in the offline RL literature, and a variety of approaches is proposed that use both model-based and model-free approaches for off-policy policy evaluation (Thomas et al. 2015, Thomas and Brunskill 2016, Le et al. 2019, Kallus and Uehara 2020). Closely related to our empirical context, Theocharous et al. (2015) use real advertising data and extend the problem of personalized ad recommendation to a dynamic setting. However, their paper only captures usage-related dynamics and ignores other dynamic ad effects such as temporal spacing, spillover, and variety. As such, the empirical results are a bit mixed with a low level of confidence in establishing gains from dynamic over myopic policies, despite their use of a high-confidence off-policy evaluation framework. Our work uses platform data with a richer state space and develops a dynamic framework that collectively

Table 1. Positioning of Our Paper with Respect to the Prior Literature on Ad Allocation

Paper	Individual-level data	Forward-looking allocation	High-dimensional state space	Dynamics-agnostic (no assumption)
Dubé et al. (2005)	✗	✓	✗	✓
Urban et al. (2013)	✓	✗	✗	✗
Wilbur et al. (2013)	✓	✓	✗	✗
Kar et al. (2015)	✓	✓	✗	✗
Schwartz et al. (2017)	✓	✗	✗	✗
Sun et al. (2017)	✓	✓	✗	✗
Theocharous et al. (2015)	✓	✓	✗	✓

incorporates dynamic effects of advertising and establishes the gains from our framework over myopic policies. More broadly, we add to the offline RL literature by presenting a model-based BIQFA algorithm and using an honest direct method that allows us to further explore the mechanism behind the gains from a dynamic policy and adds to the interpretability of our framework.

3. Setting and Data

3.1. Setting

Our data come from a leading mobile in-app advertising network of a large Asian country that had more than 85% of the market share around the time of this study. Figure 3 summarizes most key aspects of the setting. We number the arrows in Figure 3 and explain each step of the ad allocation process in detail:

- Step 1: The ad network designs an auction to sell ad slots. In our setting, the ad network runs a quasi-proportional auction with a cost-per-click payment scheme. As such, for a given ad slot and a set of participating ads $\mathcal{A}$ with a bidding profile $(b_1, b_2, \ldots, b_{|\mathcal{A}|})$, the ad slot is allocated to ad a with the following probability:

$$\pi_0(b;m) = \frac{b_a m_a}{\sum_{j\in\mathcal{A}} b_j m_j}, \tag{1}$$

where m_a is ad a's quality score, which is a measure that reflects the profitability of ad a. The ad network does not customize quality scores across auctions. The subscript 0 in π_0 refers to the fact that this is the baseline allocation policy through which our data are generated. The payment scheme is cost-per-click, similar to Google's sponsored search auctions. That is, ads are first ranked based on their product of bid and quality score, and the winning ad pays the minimum amount that guarantees their rank if a click happens on their ad.

- Step 2: Advertisers participating in the auction make the following choices: (a) design of their banner, (b) which impressions they want to target, and (c) how much to bid. Figure 3 shows an example of an auction with four different ads.

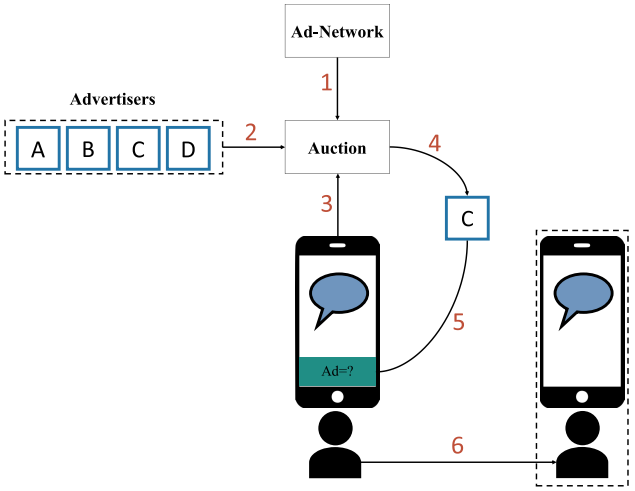
- Step 3: Whenever a user starts a new session in an app (we use a messaging app in Figure 3 as an example), a new impression is being recognized, and a request is sent to the publisher to run an auction.

- Step 4: The auction takes all the participating ads into account and selects the ad probabilistically based on the weights shown in Equation (1). All the participating ads have the chance to win the ad slot. This is in contrast with more widely used deterministic mechanisms like second-price auctions, where the ad with the highest product of bid and quality score always wins the ad slot.

- Step 5: The selected ad is placed at the bottom of the app, as shown in Figure 3.

- Step 6: Each ad exposure lasts one minute. During this time, the user makes two key decisions: (a) whether to click on the ad and (b) whether to stay in the app or leave the app and end the session. If the user clicks on the ad, the corresponding advertiser has to pay the amount determined by the auction. After one minute, if the user continues using the app, the ad network treats the continued exposure as a new impression and repeats steps 3 to 6 until the user leaves the app. We assume that a user has left the app when the time gap until the next exposure exceeds five minutes. Consistent with this definition, we define a session as the time interval between the time a user comes to an app and the time that user leaves the app.[6]

Figure 3. (Color online) Visual Schema of Our Setting

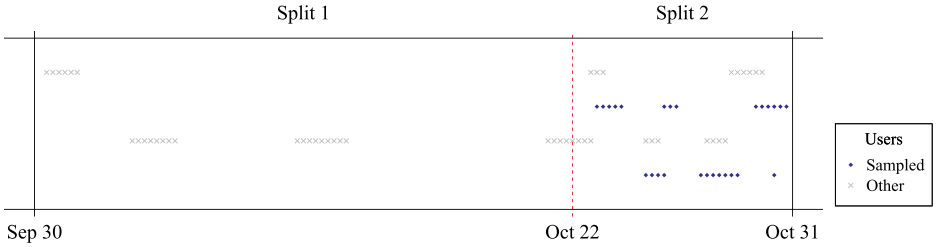


3.2. Data

We have data on all impressions and clicks for the one month from September 30, 2015, to October 30, 2015. Overall, we observe 1,594,831,699 impressions with the following raw inputs for each impression: (1) timestamp, (2) app ID, (3) user ID (Android Advertising ID), (4) GPS coordinates, (5) targeting variables that include the province, app category, hour of the day, smartphone brand, connectivity type, and mobile service provider (MSP), (6) ad ID,[7] (7) bid submitted by the winning ad, and (8) the click outcome. Importantly, our data come directly from the platform so we have access to all the information that the platform collects. Furthermore, we observe all the variables that advertisers can possibly use for targeting. Hence, we can overcome typical issues related to unobserved confounding due to the unobservability of ad assignments.

For our study, we use a sample of our full data that reflects the main goals of this paper. Because we want to optimally sequence ads within the session, our optimal intervention depends on users' history. As such, we only focus on users for whom we can use their entire history. The challenge is that no variable in our data identifies new users. As illustrated in Figure 4,

Figure 4. (Color online) Schema for Identification of New Users



our approach is to split our data into two parts based on a date (October 22) and keep users who are active in the second part of the data (October 22 to October 30), but not in the first part (September 30 to October 22). This sampling scheme guarantees that the users who are identified as new users have not had any activity in the platform at least for the three weeks prior to that. We drop all the other users from our data.

Next, we only focus on the most popular mobile app in the platform, a messaging app that has more than a 30% share of total impressions. As such, we drop new users who do not use this app. There are a few reasons why we focus on this app. First, this is the only app whose identity is known to us. Second, we expect the sequencing effects to be context dependent, so focusing on one app helps us perform a cleaner analysis. Finally, it takes users a relatively long time to learn how to use certain apps (e.g., games), and learning effects can interfere with sequencing effects. However, this messaging app is widely popular in the country and easy to use, so we expect users to pay more attention to ads from the beginning.

Overall, our sampling procedure gives us a total of 8,031,374 impressions shown to a set of 84,306 unique new users. More than 40% of these users use other apps in addition to the messaging app. In our data, there are 1,177,422 unique sessions entirely inside the focal messaging app that correspond to 6,357,389 impressions. We only focus on the impressions shown in the messaging app for our analysis. However, we use impressions shown in other apps for feature generation. Finally, it is worth noting that our sample is almost identical to that of Rafieian and Yoganarasimhan (2022b).[8] We refer the interested reader to that paper for further description of the data.

3.3. Summary Statistics

3.3.1. User-Level Variables. As discussed earlier, we sample users for whom we have the entire past history. As such, we can calculate different metrics over the entire user history and present a summary of these metrics across users. We focus on five variables and compute them using the sample of 8,031,374 impressions. We present these statistics in Table 2. We find that, on average, a user has participated in 16.23 sessions, seen 95.26 impressions and 13.97 distinct ads, and clicked 1.55 times on these impressions. Furthermore, the average CTR for a user is roughly 2%, ranging from 0% to a CTR as high as 15%. Overall, we observe a large standard deviation and a wide range for all these variables. For example, although the median number of impressions a user has seen is 40 in our data, there is a user who has seen 7,259 impressions. Thus, these statistics suggest substantial heterogeneity in user behavior that we aim to understand in our framework.

3.3.2. Distribution of Session-Level Outcomes. Our goal in this paper is to examine how much we can improve session-level user engagement through optimal sequencing of ads. As such, the key outcomes are defined at the session level. We use the sample for the focal app to compute the empirical cumulative density function (CDF) of two main outcomes of interest in this study: the total number of clicks made in a session and session length. Figure 5(a) shows the empirical CDF for the total number of clicks per session, which is our primary outcome of interest. As expected, most sessions end with no clicks on ads shown within the session, and the percentage of sessions with at least one click amounts to 6.66%. This is a reasonably high percentage in this industry. Interestingly, there are sessions with

Table 2. Summary Statistics of the User-Level Variables

Variable	Mean	Standard deviation	Minimum	Median	Maximum
Number of sessions	16.23	20.80	1	9	260
Number of impressions seen	95.26	165.62	1	40	7,256
Variety of ads seen	13.97	11.82	1	11	114
Number of clicks made	1.55	2.23	0	1	20
Click-through rate (CTR)	0.02	0.03	0	0.01	0.15

more than one click. Further exploration suggests that these sessions are typically much longer than other sessions, with an average length of more than 15 exposures.

In Figure 5(b), we show the empirical CDF of session length, as measured by the number of exposures shown within any session. This figure shows that around 50% of all sessions end in only two exposures. Furthermore, the empirical CDF in Figure 5(b) shows that the vast majority of sessions last for 10 or fewer exposures, and only a tiny fraction of them last for 30 or more exposures.

To better understand these two session-level outcomes, we focus on the outcomes at the exposure level: users' decision to click on an ad and leave the session. These decisions determine the transition dynamics of our problem. In Online Appendix A.1, we visualize the observed proportion of different transition possibility from one exposure to the next, across exposure numbers. Importantly, we find that the click decision does not necessarily lead to the leave decision.

3.3.3. Shares of Ads. Overall, we observe a total of 328 ads shown in our sample for the focal app. These ads have different shares of total impression, with some having a much higher share than others. In Online Appendix A.2, we show how each ad in our study constitutes a different fraction of total impression. In particular, we sort these ad shares in our data and demonstrate that top 15 ads account for roughly 70% of the impressions in our data. We later use this information when specifying the setting of our framework.

4. Framework for Adaptive Ad Sequencing

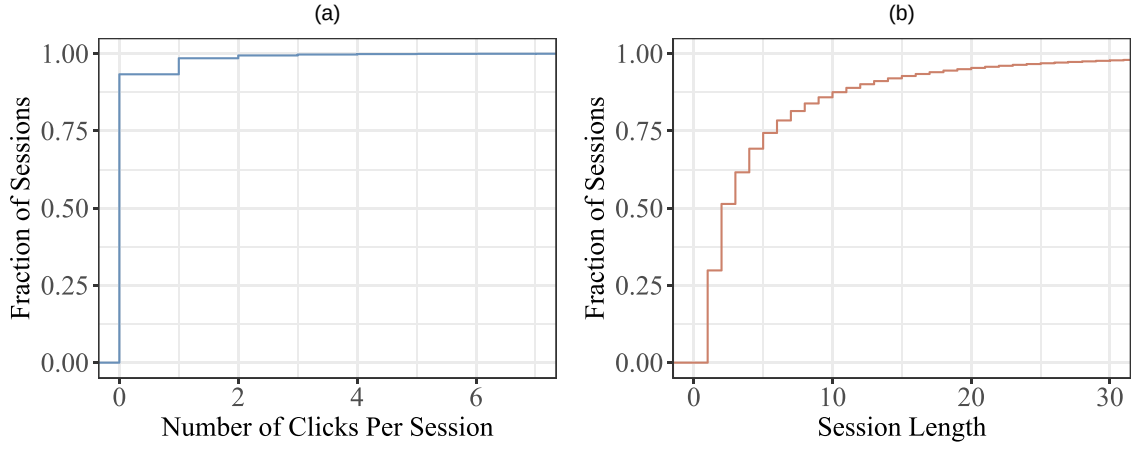
We now present our dynamic framework for the sequencing of ads. We start with the theoretical setup of our model in Section 4.1. We then use our theoretical setup to identify and address challenges in empirically designing the optimal policy in Section 4.2. Next, we discuss how we evaluate a policy using the data at hand in Section 4.3. Finally, in Section 4.4, we describe the implementation of our framework and the practical challenges that may arise.

4.1. Theoretical Setup

We begin by describing the theoretical setup of our framework. Let i denote the session, and t denote each impression in that session, for example, $t = 1$ refers to the first impression in a session. We perform our optimization at the session level, where each decision-making unit is an impression. As discussed earlier, our goal is to develop a dynamic framework that (1) captures the intertemporal tradeoffs in a publisher's ad placement decision in a session and (2) uses both presession and adaptive session-level information to personalize the sequence of ads for the user in any given session. An MDP gives us a general framework to characterize the publisher's problem and incorporate the two main goals. An MDP is a five-tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \beta \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition function, R is the reward function, and β is the discount factor. We describe each of these five elements in our context as follows:

- *State Space* ($\mathcal{S}$): The state space consists of all the information the publisher has about an exposure, which affects the publisher's decision at any time period. The publisher can take two pieces of information into account: (1) presession information and (2) session-level information. Presession information contains any data on the user up until the current session, including the user's demographic variables and behavioral history. For any session i, we denote the presession state variables by X_i. It is important to notice that the presession variables are not adaptive, that is, it does not change within the session, so we can drop the t subscript. On

Figure 5. (Color online) Empirical CDF of the Session Length and Total Number of Clicks per Session



Notes. (a) Number of clicks per session. (b) Session length (number of exposures).

the other hand, session-level variables are adaptive and change within the session. Unlike the conventional approach in MDP that restricts the state to represent only the previous time period, we consider the entire sequence of ads and users' decisions within the session. That is, for any exposure t in session i, we define $G_{i,t}$ as the set of session-level state variables as follows:

$$G_{i,t} = \langle A_{i,1}, Y_{i,1}, A_{i,2}, Y_{i,2}, \ldots, A_{i,t-1}, Y_{i,t-1} \rangle, \tag{2}$$

where $A_{i,s}$ denotes the ad shown in exposure number s and $Y_{i,s}$ denotes whether the user clicked on this ad ($s < t$). As a result, $G_{i,t}$ is the sequence of all ads and actions within the session up to the current time period. Overall, we define the state variables as $S_{i,t} = \langle X_i, G_{i,t} \rangle$, that is, a combination of both presession and session-level variables.

- *Action Space ($\mathcal{A}$)*: The action space contains the set of actions the publisher can take. In our case, this action is to show one ad from the ad inventory every time an impression is recognized. As such, $\mathcal{A}$ is the entire ad inventory in our problem.[9]
- *Transition Function (P)*: This function determines how the current state transitions to the future state, given the action made at that point. As such, we can define $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ as a stochastic function that calculates the probability $P(s' \mid s, a)$, where $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. This is a crucial component of an MDP because publishers cannot control the dynamics of the problem if the next state is not affected by the current decision. In Section 4.1.1, we discuss the components of the transition function in our problem in detail.
- *Reward Function (R)*: This function determines the reward for any action a at any state s. As such, we can define this function as $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. This function can take different forms depending on the publisher's objective. In our case, because the publisher is interested in optimizing user engagement, they can use different metrics that reflect user engagement, such as the probability that the user clicks on the ad. In Section 4.1.2, we discuss our choice of reward function in greater details.
- *Discount Factor (β)*: The rate at which the publisher discounts the expected future rewards. Given the short time horizon of the optimization problem, a risk-neutral publisher must value the current and expected future rewards equally, indicating that β is very close to one.

With all these primitives defined, we can now write the publisher's maximization problem as follows:

$$\arg\max_a [R(s,a) + \beta \mathbb{E}_{s'|s,a} V(s')], \tag{3}$$

where $V(s')$ is the value function incorporating expected future rewards at state s' if the publisher selects ads optimally. Following Bellman (1966), we can write this value function for any state $s \in \mathcal{S}$ as follows:

$$V(s) = \max_a R(s,a) + \beta \mathbb{E}_{s'|s,a} V(s'). \tag{4}$$

In summary, as shown in Equation (3), the optimization problem consists of two key elements: the current period reward and the expected future rewards. The publisher chooses the ad that maximizes the sum of these two elements.

4.1.1. Transition Function. We now characterize the law-of-motion, that is, how state variables transition given the publisher's action at any point. As mentioned earlier, we are interested in the probability of the next state being s', given that action a is taken in state s, that is, $P(s' \mid a, s)$. Suppose that the user is in state $S_{i,t} = \langle X_i, G_{i,t} \rangle$ at exposure t in session i. The only time-varying factor in $S_{i,t}$ that can transition is $G_{i,t}$, which is the history of the sequence. Given the definition of $G_{i,t}$ in Equation (2), we can determine the next state if we know the user's decision to click on the current ad and/or continue staying in the session. There are three mutually exclusive possibilities for state transitions:

- Case 1 *(click and stay):* If the user clicks on ad $A_{i,t}$ and stays in the session, we can define the next state as follows:

$$S_{i,t+1} = \langle X_i, G_{i,t}, A_{i,t}, Y_{i,t} = 1 \rangle, \tag{5}$$

where $Y_{i,t} = 1$ indicates that the user has clicked on the ad shown in exposure number t.

- Case 2 *(no click and stay):* If the user does not click on ad $A_{i,t}$ and stays in the session, we can similarly define the next state as follows:

$$S_{i,t+1} = \langle X_i, G_{i,t}, A_{i,t}, Y_{i,t} = 0 \rangle, \tag{6}$$

where $Y_{i,t} = 0$ indicates that the user has not clicked on the ad shown in exposure number t.

- Case 3 *(leave):* Regardless of user's clicking outcome, if the user decides to leave, the entire session is terminated and there is no more decision to be made. Thus, we can write

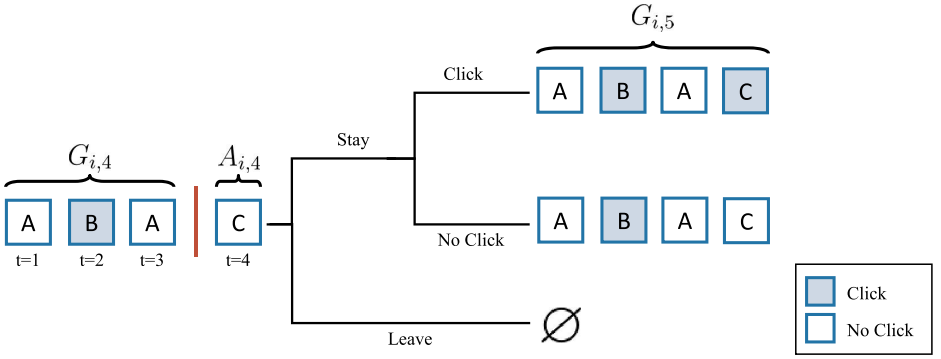
$$S_{i,t+1} = \text{ø}. \tag{7}$$

Figure 6 visually presents the three possibilities presented here. This figure illustrates an example where the publisher shows an ad in the fourth exposure in a session. It shows three possibilities and how each forms the next state. Based on this characterization, we can now define the transition function for any pair of action and state as follows:

$$P(S_{i,t+1} \mid a, S_{i,t}) = \begin{cases} (1 - P(L_{i,t} = 1 \mid a, S_{i,t})) & \\ P(Y_{i,t} = 1 \mid a, S_{i,t}) & \text{Case 1, Equation (5)} \\ (1 - P(L_{i,t} = 1 \mid a, S_{i,t})) & \\ (1 - P(Y_{i,t} = 1 \mid a, S_{i,t})) & \text{Case 2, Equation (6)} \\ P(L_{i,t} = 1 \mid a, S_{i,t}) & \text{Case 3, Equation (7)} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Equation (8) illustrates the two nondeterministic components of state transitions: click and leave probabilities. As

Figure 6. (Color online) Example Illustrating the State Transitions



such, estimating these two outcomes would be equivalent to estimating transition functions. In Section 4.2, we discuss our approach to obtain these estimates.

4.1.2. Reward Function. Another piece of an MDP that needs to be defined is the reward function. The reward function can take different forms depending on the publisher's objective. We primarily focus on maximizing the total number of clicks per session as our main objective because of a few reasons. First, clicks are the main source of revenue for the publisher because the advertiser only pays when a click happens. Second, almost all ads in our study are mobile apps whose objective is to get more clicks and installs. In the literature, this type of ad is referred to as performance ads, and their match value is generally assumed to be the probability of click (Arnosti et al. 2016). Hence, clicks are particularly good measures of user engagement with ads in our setting. Third, clicks are realized immediately in the data and well recorded without measurement error.

Given that publishers want to maximize the number of clicks made per session, we can define the reward function as the probability of click for a pair of state and action. For exposure number t in session i, we can write

$$R(S_{i,t}, a) = P(Y_{i,t} = 1 \mid a, S_{i,t}). \tag{9}$$

This is the probability of clicking on ad a if shown in the current state.

4.2. Empirical Strategy for Policy Identification

In this section, we discuss how we can take our theoretical framework to data and identify the policy that maximizes the expected rewards for each session, as characterized in our MDP. To do so, we first formally define a policy as follows.

Definition 1. A policy is a mapping $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$, that assigns a probability $\pi(a \mid s)$ to any action $a \in \mathcal{A}$ taken in any given state $s \in \mathcal{S}$.

This definition of policy allows for both deterministic and nondeterministic policies.[10] We now characterize our main goal in this section: We want to use our data to identify a policy π^* that maximizes the expected rewards for a session. That is, from the beginning to the end of a session, this policy determines which ad to show in each exposure to maximize the expected sum of rewards in that session. Following our MDP characterization, the optimal action at any given point is determined as follows:

$$\arg\max_{a \in \mathcal{A}_{i,t}} [R(S_{i,t}, a) + \beta \mathbb{E}_{S_{i,t+1} \mid S_{i,t}, a} V(S_{i,t+1})], \tag{10}$$

where $\mathcal{A}_{i,t}$ is the ad inventory, and $S_{i,t}$ is the state variable at exposure t in session i. Solving the optimization problem in Equation (10) for each possible state gives us the optimal policy function π^*.

To solve the dynamic programming problem defined in Equation (10), we face three key challenges:

- First, we need to obtain personalized estimates of the two unknown primitives in Equation (10): click and leave probabilities. That is, for any pair of state variables and ad, we need to accurately estimate the probability of click and leave. We discuss our solution to this challenge in Section 4.2.1.
- Second, our optimization is over the set of all ads. As such, even if we develop models that obtain personalized estimates of click and leave outcomes with high predictive accuracy for ads that are shown in our data, there is no guarantee that these models provide accurate estimates for the set of all possible ads (i.e., counterfactual ads). Thus, we need a framework with counterfactual validity. We describe our solution to this challenge in Section 4.2.2.
- Third, although it is, in principle, sufficient to have the estimates of reward and transition probabilities to find value functions, such an exact solution is not computationally feasible in our setting where the state space is high dimensional and grows exponentially in the number of time periods. Hence, we need an approximate solution that is scalable. We discuss our solution to this scalability issue in Section 4.2.3.

4.2.1. Personalized Estimation of Model Primitives. We start with our first challenge and formalize it as follows.

Challenge 1. *Let $\mathcal{D} = \{(S_{i,t}, A_{i,t}, Y_{i,t}, L_{i,t})\}_{i,t}$ denote the sample of impressions available, where the click and leave outcomes are recorded for each impression as $Y_{i,t}$ and $L_{i,t}$, respectively. We want to estimate functions $\hat{l}$ and $\hat{y}$ that take a pair of state variable ($S_{i,t}$) and action ($A_{i,t}$) as input and returns personalized estimates of expected click and leave probabilities as follows:*

$$\hat{y}(S_{i,t}, A_{i,t}) = \mathbb{E}(Y_{i,t} \mid S_{i,t}, A_{i,t}), \tag{11}$$

$$\hat{l}(S_{i,t}, A_{i,t}) = \mathbb{E}(L_{i,t} \mid S_{i,t}, A_{i,t}). \tag{12}$$

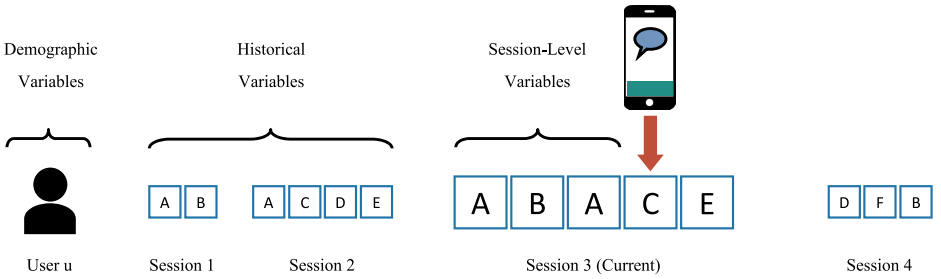
To address this challenge, we need a function that can differentiate between impressions given the available information. Because this is an outcome prediction task, we need to use machine learning methods that do not impose restrictive parametric assumptions and capture more complex relationships between the covariates and outcomes (Mullainathan and Spiess 2017). Furthermore, to allow a machine-learning algorithm to differentiate between impressions, it is essential to generate a rich set of covariates or features to represent impressions. Thus, our task becomes one of feature engineering where we want to use our domain knowledge to map $\langle S_{i,t}, A_{i,t} \rangle$ to a set of meaningful features that help us predict both click and leave outcomes.

We first define four feature categories: (1) timestamp and the ad shown in the impression that constitute the contextual information about the impression, (2) demographic features that are raw inputs about the user that are recorded by the platform, such as user's location and smartphone brand, (3) historical features that contain the information about the user's behavioral history up until the current session, such as the number of impressions the user has seen in prior sessions, and (4) session-level features that only use the information from the current session, such as the variety of previous ads shown in the session. Figure 7 provides an overview of our feature categorization. In this example, the user is at her fourth exposure in her third session. The features for this particular exposure include the observable demographic features, historical features generated from the prior sessions, and session-level features that are generated from the first three exposures shown in the current session.[11]

Our feature generation framework borrows from the literature on the advertising dynamics and behavioral mechanisms underlying these dynamics. Because the raw inputs for historical and session-level features are a user's past interactions with ads, we use features that summarize each user's long- and short-term interactions with each ad in terms of frequency akin to goodwill stock models (Nerlove and Arrow 1962, Dubé et al. 2005), recency or spacing according to memory-based models (Sawyer and Ward 1979, Sahni 2015), and clicks that have been shown to greatly help with the task of click prediction (Rafieian and Yoganarasimhan 2021). Although we use the literature to inform our feature generation, we take an agnostic approach and let our learning algorithm flexibly capture these relationships. We store these features in large inventory matrices where rows are sessions and columns are ads. This parsimonious yet rich inventory-based summarization allows us to generate other features such as ad variety and diversity as they are determined by the frequency of all ads. We further include other usage-based features such as average session length or time interval between sessions to predict the leave outcome more accurately based on the past data. Overall, our feature generation framework takes $\langle S_{i,t}, A_{i,t} \rangle$ and gives us a set of features $g(S_{i,t}, A_{i,t})$ for each impression that we can use as inputs of our learning algorithm. We present the details of all these features in Online Appendix B.

4.2.2. Counterfactual Validity. Our second challenge comes from the policy aspect of our framework: Not only do we need to obtain personalized estimates of click and leave outcomes for impressions shown in our data, but we also need to estimate these outcomes for counterfactual ads that are not shown in the data. One immediate solution is to apply our feature generation framework to counterfactual impressions and use our learning algorithm to estimate the outcomes. However, this approach can run into two key problems. First, although machine learning algorithms are known to do well in the task of interpolation, we need further guarantees on the feasibility of our counterfactual impressions for the task of extrapolation, that is, counterfactual estimation. Second, suppose the

Figure 7. (Color online) Visual Schema for Our Feature Categorization



ad assignment is confounded by an unobserved factor that is not in our feature set. In that case, the learning algorithm may incorrectly learn the link between the unobserved variable and outcomes as an ad effect. This is similar to the issue of endogeneity or selection on unobservables in the causal inference literature. We formally present these two challenges as follows.

Challenge 2. *Suppose the predictive models $\hat{y}$ and $\hat{l}$ are trained on data $\mathcal{D} = \{(S_{i,t}, A_{i,t}, Y_{i,t}, L_{i,t})\}_{i,t}$. Let $\mathcal{D}_c = \{\bigcup_{a \in \mathcal{A}_{i,t}} (S_{i,t}, a, Y_{i,t}, L_{i,t})\}_{i,t}$ denote the counterfactual data set. To ensure the counterfactual validity of our estimates on the counterfactual data, we need to address the following challenges*:

1. *For any ad $a \in \mathcal{A}_{i,t}$, the data point with the pair of state variable and action $(S_{i,t}, a)$ and the corresponding set of features $g(S_{i,t}, a)$ could have been generated in our training data $\mathcal{D}$, so finding values of $\hat{y}(S_{i,t}, a)$ and $\hat{l}(S_{i,t}, a)$ is a form of interpolation.*
2. *For any ad $a \in \mathcal{A}_{i,t}$, the assignment probability only depends on the observed set of features used in training models $\hat{y}$ and $\hat{l}$.*

To satisfy the first condition in Challenge 2, we need to identify the feasibility set $\mathcal{A}_{i,t}$ for each impression such that any ad $a \in \mathcal{A}_{i,t}$ *could have been shown* in that impression. This is equivalent to the *overlap* or *positivity* assumption in the causal inference literature that requires each treatment condition (ad in our case) to have a nonzero propensity score. That is, if $e(S_{i,t}, a)$ denotes the propensity of ad a to be shown in exposure t in session i, we must have $e(S_{i,t}, a) > 0$ for any $a \in \mathcal{A}_{i,t}$. Although attainable in principle, this is a condition that is rarely satisfied in most nonexperimental digital advertising settings because ads are selected through a deterministic allocation rule in commonly used auctions such as second-price. In our setting, however, the platform uses a quasi-proportional auction that induces randomization in ad allocation: Each ad has a nonzero propensity score if and only if it participates in an auction. As such, the propensity score is zero only when the ad is not participating in an auction due to their targeting decision or campaign availability. We use a filtering strategy similar to that in Rafieian and Yoganarasimhan (2021), where for each impression, we filter out ads that *could have never shown*. The remaining ads constitute our feasibility set $\mathcal{A}_{i,t}$, which is generally a rich set of ads given the low level of targeting in our platform. We present the details of our filtering strategy in Online Appendix C.1.

The second condition in Challenge 2 also has a strong link to the causal inference literature. Although this is a predictive task, our learning algorithm may still incorrectly learn the ad effects if there is any unobserved confounding. For example, suppose ad a_1 is more likely to be shown to less-educated adults than ad a_2, but we do not observe education in our data. Now, if less-educated adults have a higher probability of click, our learning algorithm may attribute the link between education and click to ads a_1 and a_2, if it does not control for education. Unconfoundedness is what satisfies this condition. That is, conditional on observed features $g(S_{i,t}, a)$, the assignment to ads is random. We can formally show this as a proposition in our data as follows.

Proposition 1. *In a setting with a quasi-proportional auction and observable targeting, the distribution of propensity scores is fully determined by observed covariates.*

Proof. Please see Online Appendix C.2. □

To provide empirical support for this proposition, we estimate propensity scores using observed features and assess covariate balance (see Online Appendix C.3). We then include these propensity scores $\hat{e}(S_{i,t}, a)$ in our feature set $g(S_{i,t}, a)$ to ensure that the assignment probabilities are accounted for. This further guarantees the unconfoundedness assumption as the conditional independence is satisfied only by conditioning on propensity scores (Rosenbaum and Rubin 1983).

4.2.3. Value Function Approximation.

Now, we discuss the final piece of our empirical framework to develop an optimal dynamic policy. Recall the publisher's optimization problem in Equation (10):

$$\arg\max_{a \in \mathcal{A}_{i,t}} [R(S_{i,t}, a) + \beta \mathbb{E}_{S_{i,t+1}|S_{i,t},a} V(S_{i,t+1})].$$

In Sections 4.2.1 and 4.2.2, we show how we can get the reward $R(S_{i,t}, a)$, as well as the law of motion as captured by the expectation $\mathbb{E}_{S_{i,t+1}|S_{i,t},a}$ from the previous equation. The unknown part is the value function V that captures future rewards. We can use Bellman equation to characterize this value function in a recursive relationship as follows:

$$V(S_{i,t}) = \max_{a \in \mathcal{A}_{i,t}} R(S_{i,t}, a) + \beta \mathbb{E}_{S_{i,t+1}|S_{i,t},a} V(S_{i,t+1}). \quad (13)$$

Because we know the reward function and law of motion, the typical approach to find the value function is to construct a table of all states and directly find values using Equation (13). However, this task becomes infeasible when we have a high-dimensional state space, as we need to store all the corresponding values. We can formally characterize the computational intensity of this task as follows.

Challenge 3. *Let T denote the length of the horizon over which we want to perform our optimization, and let N denote the number of sessions. For each session, our state space grows exponentially in T. Specifically, for a single session i, the order of state variables would be $O((2|\mathcal{A}_{i,1}|)^{T-1})$, because we need to record the entire ad sequence and actions (click or not click). Thus, for all sessions the complexity order would be $O((2\max_i |\mathcal{A}_{i,1}|)^{T-1} \times N)$, where $|\mathcal{A}|$ is the size of our ad inventory.*

To put things in perspective, even if we only have 10 ads in our inventory and want to perform the dynamic optimization for 10 periods, each session has the complexity order of 10^9. Now, if we want to that for the number of sessions in our data that is roughly one million, the order of complexity would be 10^{15}. As such, conventional tabular solutions in the marketing and economics literature cannot work in our problem.

To address this challenge, we turn to the literature on value function approximation in dynamic programming and RL (Sutton and Barto 2018). Our solution is to develop a function approximation algorithm that approximates the value function instead of finding all the values directly. That is, we want to learn a function $\hat{v}: \mathcal{S} \rightarrow \mathbb{R}$ with a set of parameters θ_v. This approach can significantly reduce the time complexity because we need only an order of magnitude smaller subset of states to learn a function, and the representation of this function is only through the set of parameters θ_v.

Before we present our algorithm, we first introduce a new notation. We define a function $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ to represent the entire term that the publisher maximizes in Equation (10) as follows:

$$Q(S_{i,t}, a) = R(S_{i,t}, a) + \beta \mathbb{E}_{S_{i,t+1}|S_{i,t},a} V(S_{i,t+1}). \tag{14}$$

The Q function is often referred to as the choice-specific value function in the econometrics literature (Aguirregabiria and Mira 2002). Given the Bellman equation in Equation (13), we can write

$$Q(S_{i,t}, a) = R(S_{i,t}, a) + \beta \mathbb{E}_{S_{i,t+1}|S_{i,t},a} \max_{a' \in \mathcal{A}_{i,t+1}} Q(S_{i,t+1}, a'). \tag{15}$$

Now, we can use our transition function in Equation (8) and plug in our estimates for click and leave probabilities to define $\tilde{Q}_t$ in a similar way to Equation (15) as follows:

$$\begin{aligned} \tilde{Q}_t(S_{i,t}, a) = {} & \hat{y}(S_{i,t}, a) + (1 - \hat{l}(S_{i,t}, a))\hat{y}(S_{i,t}, a) \\ & \max_{a' \in \mathcal{A}_{i,t+1}} \tilde{Q}_{t+1}(\langle S_{i,t}, a, Y_{i,t} = 1 \rangle, a') \\ & + (1 - \hat{l}(S_{i,t}, a))(1 - \hat{y}(S_{i,t}, a)) \\ & \max_{a' \in \mathcal{A}_{i,t+1}} \tilde{Q}_{t+1}(\langle S_{i,t}, a, Y_{i,t} = 0 \rangle, a'), \end{aligned} \tag{16}$$

where the first term $\hat{y}(S_{i,t}, a)$ is the current period reward, and the other two elements in the right-hand side (RHS) of Equation (16) capture the two transition possibilities where the session still continues: "click and stay" and "no click and stay".

Function $\tilde{Q}_t$ represents a plugin version of our Q function in Equation (14) at time period t, where we directly plug in our reward and transition estimates to find the Q values.[12] Our goal is to estimate a function $\hat{q}_t$ that approximates $\tilde{Q}_t$. However, this task is not trivial as these functions appear in both the left-hand side (LHS) and RHS of Equation (16). We can follow the common insight in the literature to formulate an iterative procedure such as value iteration or backward induction to simplify the task to supervised learning. In our framework, we focus on backward induction as it is reasonable to assume a finite horizon because most sessions end in a few exposures. Furthermore, for a short length of horizon T, the backward induction algorithm runs faster than a value iteration algorithm because value iteration may require far more iterations for convergence.

The logic behind backward induction for q-function approximation (BIQFA) is simple: From the set of $\{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T\}$, we learn the functions one at a time in a backward order. We start with the last time period T where the function $\hat{q}_T$ is equivalent to our click prediction function $\hat{y}$ because this is the last period and the future rewards are assumed to be zero.[13] We can then complete the RHS of Equation (16) and obtain the plugin outcomes for any subset of states in period $T-1$. These plugin outcomes are often referred to as Bellman backups and denoted by $\tilde{Q}$ (Lee et al. 2021). Once we have these plugin outcomes, the task of estimating $\hat{q}_{T-1}$ simplifies to one of supervised learning, where we can use our set of state variables and actions to estimate the plugin outcomes or Bellman backups. We can continue this process until we have the full set of functions $\{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T\}$.

Before we present our algorithm in detail, we define the set of inputs and outputs of the algorithm. Let $\tilde{S}_t$ denote a subsample of the state space at exposure t. The algorithm takes data $\mathcal{D}$, functions of click, leave, and propensity score estimates $(\hat{y}, \hat{l}, \hat{e})$, length of horizon (T), and subsamples of the full state space at each exposure ($\tilde{S}_t$ for all $t \leq T$) as inputs, and return the set of q-functions ($\hat{q}_t$ for all $t \leq T$) as outputs. Our BIQFA algorithm is presented in detail in Algorithm 1.

Algorithm 1 (BIQFA)

Input: $\mathcal{D}, \hat{y}, \hat{l}, \hat{e}, T, \tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_T$ ▷ $\tilde{S}_t \subset \mathcal{S}$ at exposure t
Output: $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T$

$\hat{q}_T \leftarrow \hat{y}$
for $t = T-1 \rightarrow 1$ **do**
  $\tilde{Q}_{t+1} \leftarrow \hat{q}_{t+1}$
  for each $s \in \tilde{S}_t, a \in \mathcal{A}$ **do**
    $\bar{Q}_{s,a} \leftarrow \tilde{Q}_t(s,a)$ ▷ Create Bellman backups using Equation (16)
    if $\hat{e}(s,a) = 0$ **then**
      $\bar{Q}_{s,a} = 0$
    end if
    $Z_{s,a} \leftarrow \{g(s,a), \hat{y}(s,a), \hat{l}(s,a)\}$ ▷ Set of inputs given to the learning algorithm
  end for
  $\hat{q}_t \leftarrow$ **learn**$(Z_{s,a}, \bar{Q}_{s,a})$ ▷ Any learning algorithm can be used
end for

A few details are worth noting about our BIQFA algorithm. First, the time-saving component of our approximation framework is in sampling $\tilde{S}_t$ from the

full state space $\mathcal{S}$. As such, we want $|\tilde{S}_t|$ to be not very large but representative of states that would be generated under the optimal dynamic policy, so the algorithm can learn a good approximation of the q-function at a reasonable computational cost.[14] However, the challenge is that we do not know the distribution of states under the optimal dynamic policy before running the algorithm. Therefore, we need a good initialization that is close to the distribution of states under the optimal dynamic policy. A good candidate is to use an adaptive myopic policy that selects the ad with the highest reward at any point (i.e., $\arg\max_{a\in\mathcal{A}_{i,t}}\hat{y}(S_{i,t},a)$ for any state variable $S_{i,t}$), which is a special case of optimal dynamic policy when $\beta = 0$. As a result, the distribution of this policy is likely close to that of optimal dynamic policy so we use a sample of states under the adaptive myopic policy for initialization.[15] The exact size of each $|\tilde{S}_t|$ can be set a priori by the researcher or through a validation procedure described in Online Appendix D.1. Second, although our set of generated features $g(s, a)$ suffices in principle for learning q-functions, we include click and leave predictions as features to help the learning algorithm capture the dynamic structure more easily. As such, the specific input of $\hat{q}$ functions is $Z_{s,a}$, which contains the generated features and click and leave estimates (line 9 of our algorithm). Third, given that we use propensity scores in our feature set $Z_{s,a}$, the learning algorithm easily learns the association between zero propensity and zero Bellman backup.

Last, we discuss the convergence properties of our BIQFA algorithm. The idea of value function approximation has been around since Samuel (1959) and Bellman and Dreyfus (1959), and many algorithms have been proposed for this task to this date with significant practical success (Mnih et al. 2015, Sutton and Barto 2018). The early theoretical studies on the convergence properties of function approximation are Gordon (1995) and Tsitsiklis and Van Roy (1996), who show under what conditions we have convergence. The main issue is that most of these requirements for convergence are violated when we use more high-capacity learners such as deep learning or XGBoost, and it is easy to show divergence using counterexamples (Levine et al. 2020). However, some recent studies show that using these high-capacity function approximators generally tend to converge in practice, as they correspond to a very large class of functions (Van Hasselt et al. 2018, Fu et al. 2019). In the absence of theoretical convergence guarantees on our algorithm, we present some results in Online Appendix D.2 to establish its strong performance in our data.

In sum, our BIQFA algorithm approximates the set of $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T$ needed to identify the adaptive ad sequencing policy. It is important to notice that like other function approximation methods in the literature, the computational complexity of our BIQFA algorithm is not exponential. Increasing the length of horizon only increases the computational complexity of our algorithm linearly as we need to approximate a higher number of $\hat{q}_t$. Similarly, increasing the number of ads increase the computational complexity polynomially, because it changes the number of observations in for each t (line 4 of the algorithm), and the dimensionality of $Z_{s,a}$. Therefore, BIQFA is scalable to large T and number of ads. Our BIQFA algorithm differs from the conventional approaches in the RL literature such as fitted Q-iteration (FQI) in two ways. First, our approach is model based; that is, our algorithm uses the model-based estimates of the transition function. We use this approach because there are probabilistic components in state transitions in our problem that have low probability of occurring such as clicks, so a model-free approach would not perform very well in these domain. Our model-based estimates of the transition stabilizes the function approximation procedure. Second, as discussed earlier, we use a backward induction solution concept as opposed value iteration. This choice allows us to obtain a function approximation in fewer iterations.

4.3. Evaluation

Once we identified the optimal dynamic policy for adaptive ad sequencing using our empirical framework in Section 4.2, we need to evaluate this policy and compare it to other benchmarks. As such, we need an evaluation framework that takes any policy π^* and data $\mathcal{D}$ as input and evaluates the policy in terms of the outcomes of interest, specifically the expected number of clicks per session. This task is often referred to as *counterfactual policy evaluation* in the marketing and economics literature and *off-policy policy evaluation* in the RL literature.

The fundamental problem is that the data at hand are often generated by a *behavior policy* π^b, which is different from the policies we want to evaluate (π^*). In a case like that, there are many approaches to evaluate the policy π^*. The common approach in marketing and economics literature is to use a counterfactual simulation approach, where we simulate the data given policy π, using the estimates for reward and transition functions (Dubé et al. 2005, Simester et al. 2006). This approach is often referred to as the *direct method* (DM) in the RL literature as it directly uses model estimates to evaluate the policy (Kallus and Uehara 2020). An important advantage of this approach is that it can capture the heterogeneity at the most granular level, which is session level in our case. That is, we can evaluate each session under a policy and examine which sessions have higher gains. On the other hand, the main issue with the DM is that reward and transition estimates may be largely biased in the absence of randomization, which results in a biased policy evaluation. In our setting, we have randomization in ad allocation that satisfies the unconfoundedness assumption.

Thus, the typical challenges with the DM approach are not present in our setting.

Nevertheless, there is still an important challenge in DM when it comes to policy evaluation.

Challenge 4. *Let $\mathcal{D}_{Model}$ denote the data used for policy identification, and $\mathcal{D}_{Evaluation}$ denote the data used for policy evaluation. If $\mathcal{D}_{Model} = \mathcal{D}_{Evaluation}$, then our evaluation always shows a better performance for the identified optimal dynamic policy, because our policy identification framework chooses a policy if it is best-performing given $\mathcal{D}_{Model}$ and models trained on it.*

This is an important theoretical issue, which is often unaddressed in counterfactual policy evaluation in the structural econometrics literature. To ensure that our imposed structure does not force a certain outcome, we follow the insights from the evaluation approach in Mannor et al. (2007) and double q-learning in Hasselt (2010) for de-biasing the value function estimates through sample splitting such that $\mathcal{D}_{Model} \cap \mathcal{D}_{Evaluation} = \varnothing$. We call this approach *honest direct method* (HDM) and present it in the step-by-step procedure as follows:

- Step 1: We split the data into three parts: $\mathcal{D}_{Model}$, $\mathcal{D}_{Evaluation}$, and $\mathcal{D}_{Test}$.
- Step 2: We use our modeling data $\mathcal{D}_{Model}$ to estimate functions needed for policy identification: $\hat{y}^M$, $\hat{l}^M$, and $\hat{q}_t^M$ for any t (notice that superscript M refers to the data used for estimation). We can use these functions to identify the optimal policy π^M.
- Step 3: We use our evaluation data to estimate the model primitives: probability of click and leave. The functional estimates of these primitives are denoted by $\hat{y}^E$ and $\hat{l}^E$. We use these estimates to simulate the data under any counterfactual policy, where superscript E refers to the fact that we use the evaluation data.
- Step 4: For any session in $\mathcal{D}_{Test}$, we use our policy π^M from Step 2 and our estimates $\hat{y}^E$ and $\hat{l}^E$ from Step 3 to simulate the data under the policy and evaluate its outcomes. Although we can run large-scale simulations to evaluate the outcome, there is an analytical derivation for our HDM. For any exposure t, let g_t denote a t-step trajectory of states, actions, and rewards as follows:

$$g_t = \langle s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t, a_t, r_t \rangle, \tag{17}$$

where s, a, and r denote state, action (ad), and the reward outcome, respectively. The probability of any arbitrary g_t is determined by the policy π^M and transition functions $\hat{y}^E$ and $\hat{l}^E$. For brevity, we use γ^E to denote the joint distribution of transitions. The trajectory g_t comes from the joint distribution (π^M, γ^E), where the policy comes from Step 2 and transitions come from Step 3 to satisfy our honesty criteria, which means that the data and models used for policy identification are different from those used for policy evaluation.

Now, for any session i with initial state $S_{i,1}$ and policy π^M, we can define the policy evaluation function ρ as follows:

$$\rho(\pi^M; S_{i,1}, T) = \mathbb{E}_{g_t \sim (\pi^M, \gamma^E)} \left[\sum_{t=1}^{T} \beta^{t-1} r_t \middle| s_1 = S_{i,1} \right], \tag{18}$$

where T denotes the horizon length and the expectation is taken over all trajectories. Although all trajectories constitute a massively large set, we can develop different algorithms to perform this task more efficiently and find $\rho(\pi^M; S_{i,1}, T)$. We describe the algorithm we use in Online Appendix E.1.

Overall, by splitting our data into three sets, our HDM approach overcomes two important issues with a model-based evaluation: (1) using a separate test set to perform policy evaluation avoids the issues of overfitting, and (2) separating the modeling and evaluation data sets ensures that the imposed structure of policy evaluation does not systematically favor one policy over another. That is, any other policy can theoretically outperform our optimal dynamic policy. Finally, it is worth noting that with a large T, this exact evaluation procedure can become computationally intensive for any dynamic policy. A simple solution in these cases is to simulate a few instances for each session and take the average outcome.

4.4. Practical Considerations and Implementation

Although our framework is set up more generally to be broadly applicable to other domains, there are many elements that we need to set given the context, such as the length of the horizon or the size of action space (ad inventory). We discuss these practical details in this section as follows:

- First, we need to set the length of horizon T. From our data, we observe that more than 85% of sessions end in 10 or fewer exposures (Figure 5(b)). As such, $T = 10$ is a reasonable choice as the majority of events happen in the first ten exposures. However, it is worth emphasizing that the computational complexity increases only linearly in T in our function approximation framework.
- Second, we need to define the ad inventory. An obvious choice would be to focus on our inventory's entire set of ads. Although our framework is computationally scalable to having a large action space, it would be practically difficult to obtain accurate, personalized estimates for ads with low frequency in data. As a result, we only focus on the top 15 ads with the highest frequency in our data that collectively generate over 70% of all impressions.[16]
- Third, we need to set a splitting rule for $\mathcal{D}_{Model}$, $\mathcal{D}_{Evaluation}$, and $\mathcal{D}_{Test}$. We split our data at the user level according to an approximately 40-40-20% rule such that $\mathcal{D}_{Test}$ contains sessions for 20% of users and $\mathcal{D}_{Model}$ and $\mathcal{D}_{Evaluation}$ each represents 40% of users. The specific details of our splitting procedure is presented in Online Appendix E.2.

• Fourth, we need to choose a learning algorithm and a validation procedure for the task of estimating click and leave outcomes, that is, functions $\hat{y}^M$, $\hat{l}^M$, $\hat{y}^E$, and $\hat{l}^E$. Generally, one could use any learning algorithm to estimate these functions. In our study, we use the XGBoost method developed by Chen and Guestrin (2016), which is a fast and scalable version of boosted regression trees (Friedman 2001). There are some key reasons why we use XGBoost as our main learning. First, it has been shown to outperform most existing methods in most prediction contests, especially those related to human decision-making like ours (Chen and Guestrin 2016). Second, Rafieian and Yoganarasimhan (2021) show that in the same context, XGBoost achieves the highest predictive accuracy compared with other methods. Following the arguments in Rafieian and Yoganarasimhan (2021), we use the logarithmic loss as our loss function. To tune the parameters of XGBoost, we use a hold-out validation procedure to prevent the model from overfitting. We select the hyper-parameters accurately using a grid search over a large set of hyper-parameters and select those that give us the best performance on a hold-out validation set. For more details, please see Online Appendix F.

• Fifth, for the task of q-function approximation in our BIQFA algorithm, we need to specify a learning algorithm. For internal consistency, we use XGBoost as our learning algorithm.

In sum, the choices above are made not because of the limitations in our framework but rather according to the specifics of our context. In a different context, one may need to change these decisions to get the best out of this framework. It is worth noting that our empirical application does not face the cold-start problem. We separately discuss the robustness of our framework to the cold-start problem in Online Appendix G by presenting solutions to address the problem in Online Appendices G.1 and G.2.

5. Results

5.1. Predictive Accuracy of Machine Learning Models

In this section, we examine the predictive accuracy of our click and leave estimation models. We focus on two different metrics that capture different aspects of the predictive performance:

• *Relative Information Gain (RIG):* This metric reflects the percentage improvement in logarithmic loss compared with a baseline model that simply predicts the average CTR for all impressions. We use *RIG* as our primary metric as it is defined based on the log loss, which is the loss function we used in our XGBoost models to estimate click and leave outcomes.

• *Area Under the Curve (AUC):* It determines how well we can identify *true positives* without identifying *false positives*. This score ranges from zero to one, and a higher score indicates better performance and greater classification.

These two metrics are commonly used to evaluate the predictive performance of click prediction models. In general, *RIG* is more relevant when we want to evaluate how well our model estimates the probabilities, whereas *AUC* demonstrates how good a classifier our model is. For both metrics, a higher value means better performance.

We now evaluate the predictive performance of our click and leave estimation models. As discussed earlier in Section 4.3, our honest direct method estimates two separate models for each outcome: one using the modeling data $\mathcal{D}_{Model}$ and the other using the evaluation data $\mathcal{D}_{Evaluation}$. This gives us a total of four models $\hat{y}^M$, $\hat{y}^E$, $\hat{l}^M$, and $\hat{l}^E$. We present both *RIG* and *AUC* for each of these models when evaluated on modeling, evaluation, and test samples separately.

We present our results in Table 3. In the top two panels, we examine the predictive accuracy of our click models. The model achieves a >0.20 *RIG* on the test set, which demonstrates a substantial predictive accuracy compared with the literature (Yi et al. 2013). Furthermore, both in- and out-of-sample, our click models achieve an *AUC* of more than 0.80, which shows a good classification performance by the model.

The last two panels in Table 3 show how our leave models perform. Unlike our click models, we do not expect our leave model to reach a very high predictive accuracy because app usage is less dependent on ad exposures and more driven by the app. This is particularly challenging for a messenger app where users' decision to leave primarily stems from their messaging behavior, which is unobserved to the advertising platform. Despite these limitations, both our *RIG* and *AUC* measures show information gain from our predictive model compared with average estimators. Thus, our approach to endogenize usage is advantageous over a bulk of papers in the literature that rely on simple average estimates for continuation probabilities (Kempe and Mahdian 2008, Kar et al. 2015, Sun et al. 2017).[17] In Online Appendix H, we further explore the contribution of different types of features to our predictive models.

5.2. Counterfactual Policy Evaluation

We now use our HDM to evaluate the performance of our adaptive ad sequencing framework and compare it to competing benchmarks. We refer to the policy developed by our framework as *fully dynamic* or *adaptive forward-looking* interchangeably throughout. We now define a series of competing policies for benchmarking[18]:

• *Adaptive Myopic Policy:* This policy uses all the information available at any exposure and selects the ad that maximizes the reward at that point, that is, the

Table 3. Predictive Accuracy of XGBoost Models for Click and Leave Estimation

Model	Outcome	Training Sample	Metric	Sample $\mathcal{D}_{Model}$	Sample $\mathcal{D}_{Evaluation}$	Sample $\mathcal{D}_{Test}$
$\hat{y}^M$	Click	$\mathcal{D}_{Model}$	*RIG*	0.2123	0.1988	0.2021
			AUC	0.8229	0.8110	0.8139
$\hat{y}^E$	Click	$\mathcal{D}_{Evaluation}$	*RIG*	0.2019	0.2175	0.2024
			AUC	0.8138	0.8283	0.8138
$\hat{l}^M$	Leave	$\mathcal{D}_{Model}$	*RIG*	0.1009	0.0882	0.0881
			AUC	0.7189	0.7055	0.7047
$\hat{l}^E$	Leave	$\mathcal{D}_{Evaluation}$	*RIG*	0.0880	0.1005	0.0877
			AUC	0.7051	0.7188	0.7045

highest CTR. This policy is myopic as it ignores the expected future rewards and is equivalent to $\beta = 0$ in our MDP in Equation (10). However, this policy is adaptive because it uses the real-time updated session-level features as it moves forward. We use this policy as the main comparison point for our framework because it reflects the standard practice in the advertising industry, where the platforms use a version of contextual bandit to select the ad at any point (Theocharous et al. 2015).

- *Single-Ad Policy:* This policy selects a single ad to show for the entire session. As such, this policy is not adaptive as it only uses presession information (demographic and historical features) to select the ad with the highest CTR. Using this policy as a benchmark is important from a managerial standpoint because it mimics the practice of using a fixed ad slot as opposed to a refreshable ad slot. Furthermore, it highlights the value of adaptive session-level information.
- *Random Policy:* This policy randomly selects an ad from the ad inventory at any point. Although this is a naïve policy, it is often used in the RL literature as a benchmark.

We document the performance of our *fully dynamic* policy and these three benchmarks in terms of different outcomes in Table 4. We start with the main metric of interest in this paper: the expected number of clicks per session. This metric determines how many clicks each policy generates in total when we multiply it by the number of sessions. Our results in the first row of Table 4 show that the *fully dynamic* policy developed by our adaptive ad sequencing framework results in substantial gains by achieving an expected number of 0.1671 clicks per session. In particular, the *fully dynamic* policy generates 5.76%, 27.46%, and 79.59% more clicks than *adaptive myopic*, *single-ad*, and *random* policies, respectively. The gains from our *fully dynamic* policy over the *single-ad* policy illustrates the opportunity cost of using a nonrefreshable ad slot that only shows one ad for the entire session. More importantly, the gains from our *fully dynamic* policy over the *adaptive myopic* policy make a compelling case for the use of dynamic optimization and RL in the advertising domain and call for a change in the current practice of using myopic frameworks, particularly in cases like ours where users are exposed to multiple ads sequentially over a short period of time.

Next, we aim to identify the primary source for the gains from the *fully dynamic* policy. As discussed earlier, there are two channels through which adaptive ad sequencing can create value: (1) by making users stay longer, thereby increasing the total number of impressions generated (extensive margin), or (2) by making each impression more likely to receive a click (intensive margin). We test each source using two other metrics: expected CTR for each impression and expected session length. We find that each impression has a significantly higher probability of receiving a click, but the increase in usage is only 0.2% compared with the *adaptive myopic* policy. Thus, adaptive ad sequencing increases the total number of clicks in a session by increasing the response rate to each ad. Later, in Section 5.3, we further explore the mechanism behind the increase in response rate through sequencing.

Finally, we examine how concentrated the ad allocation is under each policy. We first calculate the average share of each ad under each policy and then use the well-known HHI to measure ad concentration. Lower HHI values indicate a lower ad concentration and more evenly distributed shares. Naturally, we expect the random sequencing policy to have a very low HHI as it most evenly distributes ad shares. We observe that in the fifth row of our table. Interestingly,

Table 4. Performance of Different Sequencing Policies in the Test Data

Metric	Sequencing policies: Fully dynamic	Adaptive myopic	Single-ad	Random
Expected no. of clicks per session	0.1671	0.1580	0.1311	0.0930
Percentage click increase over random	79.59%	69.81%	40.90%	0.00%
Expected CTR (per impression)	4.26%	4.04%	3.43%	2.42%
Expected session length	3.9258	3.9164	3.8246	3.8518
Ad concentration (HHI)	0.2902	0.3178	0.3480	0.1159
No. of users	14,084	14,084	14,084	14,084
No. of sessions	201,466	201,466	201,466	201,466

we find that our *fully dynamic* policy results in a lower HHI than both *adaptive myopic* and *single-ad* policies. This is likely because the dynamic policy makes better use of synergies between ads, thereby increasing the shares for less popular ads. This is an important finding because it means that the better performance of the *fully dynamic* policy does not come at the expense of less popular ads. The lower ad concentration can also have welfare impacts for consumers as they are exposed to a more diverse set of ads.

5.3. Interpretation and Mechanism Analysis

In principle, all the differences between the *fully dynamic* and *adaptive myopic* policies stem from the fact that only the former takes into account the expected future rewards when making a decision. As such, we expect the *fully dynamic* policy to perform better than the *adaptive myopic* policy in later exposures within the session. Figure 8 confirms this pattern by breaking down the expected rewards for both policies (Figure 8(a)) and the gains from the *fully dynamic* policy over *adaptive myopic* policy across exposure numbers (Figure 8(b)). Although the *fully dynamic* policy performs worse than the *adaptive myopic* policy in the first two exposures, the gains from the *fully dynamic* policy appear from the third exposure onward. The existence of this pattern further highlights the value of scalability in our framework that allows us to extract value from the later exposures.

In summary, the observed difference in Figure 8 is because of the intertemporal tradeoff the *fully dynamic* policy makes as captured by expected future rewards in Equation (10), that is, $\beta \mathbb{E}_{S_{i,t+1}|S_{i,t},a} V(S_{i,t+1})$. Although this additional term in the equation helps achieve a better performance, interpretation of it is generally very hard as many factors go into the construction of value function. Our main goal in this section is to use the domain knowledge in advertising to add to the interpretability of our framework and share insights into the possible mechanisms behind the gains from it. First, in Section 5.3.1, we demonstrate the heterogeneity in gains from using a *fully dynamic* policy over an *adaptive myopic* policy and find the correlates of these session-level gains using the historical features for each session. We then quantify the extent of difference between the two policies and show how the historical features help explain these discrepancies in Section 5.3.2.

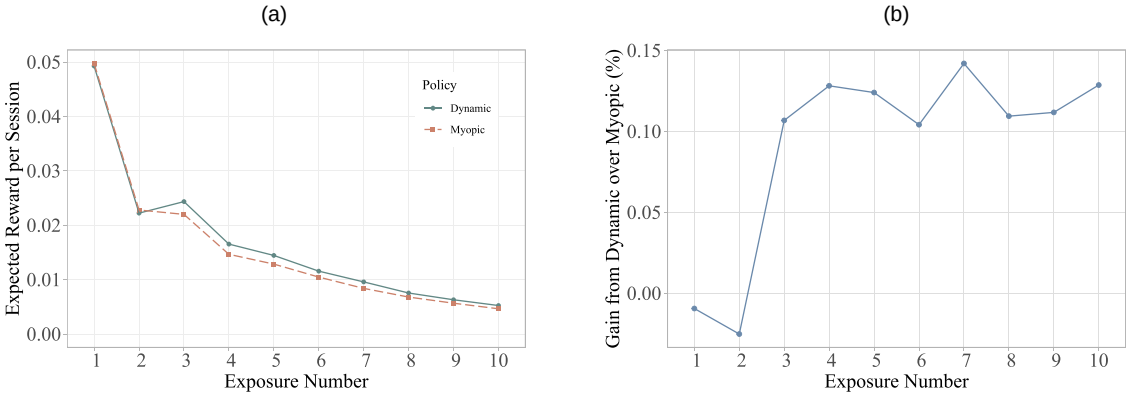
5.3.1. Heterogeneity in Gains Across Past Historical Features. In this section, we want to better understand the heterogeneity in gains from *fully dynamic* over *adaptive myopic* policy. As such, we need to first formally define gains. Let π_d^M and π_m^M denote the *fully dynamic* and *adaptive myopic* policies identified using the modeling data $\mathcal{D}_{Model}$. For each session i, we use Equation (18) to define the variable $Gain_i$ as follows:

$$Gain_i = \frac{\hat{\rho}(\pi_d^M; S_{i,1}, T = 10)}{\hat{\rho}(\pi_m^M; S_{i,1}, T = 10)} - 1, \tag{19}$$

where $\hat{\rho}(\pi_d^M; S_{i,1}, T = 10)$ and $\hat{\rho}(\pi_m^M; S_{i,1}, T = 10)$ represent the expected number of clicks for the first 10 exposures of session i with initial state variables $S_{i,1}$, under *fully dynamic* and *adaptive myopic* policies, respectively. The variable $Gain_i$ measures the percentage improvement in expected rewards from the *fully dynamic* over *adaptive myopic* policy for any specific session. Thus, it allows us to document the heterogeneity in gains across sessions.

We first focus on a simple variable that is available prior to any session: the number of previous sessions the user has experienced. We want to see how the gains change with the number of prior sessions. On the one hand, we know that a richer history helps learn user preferences more accurately. This can favor the *fully dynamic* policy and increase gains because this policy will use more accurate predictions about session dynamics (e.g., how long the session will last). On the other hand, more experience comes with a

Figure 8. (Color online) Performance of Fully Dynamic and Adaptive Myopic Policies Across Exposure Numbers

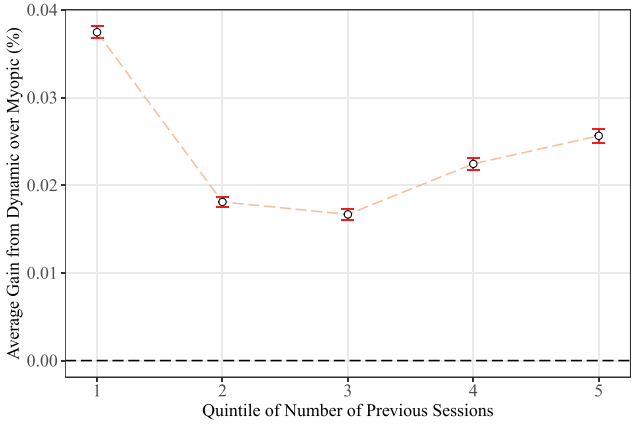


Notes. (a) Expected reward over exposure number. (b) Percentage gain over exposure number.

higher variety of prior ads, which reveals more about ad-specific user preferences. This increase in predictive accuracy can make the two policies more similar, thereby reducing gains. Furthermore, we know a higher variety of ads in the past reduces the novelty of ad interventions in the present, which can reduce the effectiveness of sequencing strategies. For example, Rafieian and Yoganarasimhan (2022b) show that as users become more experienced, the impact of an increase in variety decreases. We want to show how the mix of the opposing forces described previously would shape the overall relationship between the number of prior sessions and gains. For all the sessions in our test data, we define five quintiles based on the number of prior sessions.[19] We show the average gains for each quintile in Figure 9. Interestingly, we find a U-shaped pattern consistent with the opposing accounts presented above. Later in Section 5.4, we show that the U-shaped pattern is robust to alternative specifications.

Inspired by the pattern in Figure 9, we further document the heterogeneity in gains across a richer set of historical covariates. We first include two historical features that are correlated with the number of prior sessions and correspond to the opposing accounts presented previously: (1) the number of past impressions and (2) the variety of ads seen. We expect a positive association in the former as it demonstrates the amount of data available, whereas a negative association in the latter as a higher variety of ads seen likely makes the policies more similar and users less responsive to sequencing. We regress gains for each session on these two covariates while controlling for user and hour fixed effects to ensure our estimates do not capture user-level differences and supply-side factors such as advertisers' targeting and availability. We exclude the first session for each user because the historical features do not exist for those sessions. We present the results of this model in the first column of Table 5. Our results provide further support for the two accounts presented earlier. The coefficient for the *number of past impressions* is positive, whereas the coefficient for the *variety of ads seen* is negative. Together, having seen more impressions increases the gains, whereas having seen more distinct ads decreases the gains.

Figure 9. (Color online) Average Gains from Dynamic Policy over Myopic Policy Across Quintiles for the Number of Past Sessions



In columns (2)–(4), we add historical features one by one. We first add the *number of past clicks* prior to the current session. A higher value of this covariate indicates a greater ad response and overall engagement with ads. As shown in the second column of Table 5, this covariate has a positive association with gains from ad sequencing. Next, we include another historical feature: the *time since the last session* (in hours). Higher values of this covariate show lower recency in users' interaction with ads. In general, we expect higher recency to reduce the novelty of ad interventions, thereby lowering the gains from ad sequencing. We confirm this prediction by finding a positive coefficient for the *time since the last session* in column (3) of Table 5: The greater the gap is between the current session and the last session, the higher the gains are from sequencing. Finally, we include the *length of the last session* as another covariate in our model. This covariate is a signal for the length of the current session. As shown in Figure 8, gains from sequencing appear later in a session. Hence, when the session is longer, we expect the gains to be higher. The positive and statistically significant coefficient for the *length of last session* in the fourth column of Table 5 provides support for this prediction.

5.3.2. Extent of Discrepancy Between Dynamic and Myopic Policies. The key takeaway from the previous section is that there is great heterogeneity in gains from sequencing across past historical features. These gains naturally stem from the differences between the *fully dynamic* and *adaptive myopic* policies. In this section, we want to see where the discrepancy between the two policies is more pronounced. As such, we first need to quantify the discrepancy between the two policies at the session level. For any given session i and policy π, we can determine the distribution of ad shares both analytically and through simulations. Let $\boldsymbol{\alpha}_i^{(d)}$ and $\boldsymbol{\alpha}_i^{(m)}$ denote vectors representing ad shares in session i under *fully dynamic* and *adaptive myopic* policies, respectively. We quantify the discrepancy between these two distributions using five measures based on ℓ-norm and Kullback-Leibler (KL) divergence as follows:

- Outcome 1: ℓ^1-norm of the difference between shares $\|\boldsymbol{\alpha}_i^{(d)} - \boldsymbol{\alpha}_i^{(m)}\|_1$
- Outcome 2: ℓ^2-norm of the difference between shares $\|\boldsymbol{\alpha}_i^{(d)} - \boldsymbol{\alpha}_i^{(m)}\|_2$

Table 5. Heterogeneity in Gains from Dynamic Policy over Myopic Policy Across the Historical Features

Historical features	Dependent variable: $Gain_i$			
	(1)	(2)	(3)	(4)
No. of past impressions	0.00001***	0.00001**	0.00001***	0.00001*
	(4.60)	(3.00)	(3.40)	(2.07)
Variety of ads seen	−0.00024*	−0.00028**	−0.00021*	−0.00033**
	(−2.43)	(−2.79)	(−2.10)	(−3.28)
No. of past clicks		0.00076***	0.00080***	0.00080***
		(3.70)	(3.89)	(3.91)
Time since last session			0.00008***	0.00007**
			(4.84)	(4.17)
Last session length				0.00036***
				(22.23)
User fixed effects	✓	✓	✓	✓
Hour fixed effects	✓	✓	✓	✓
No. of observations	190,206	190,206	190,206	190,206
R^2	0.271	0.271	0.271	0.273
Adjusted R^2	0.220	0.220	0.220	0.222

Note. Numbers in parentheses are t statistics that are estimated using OLS.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

- Outcome 3: KL divergence of $\boldsymbol{\alpha}_i^{(d)}$ from $\boldsymbol{\alpha}_i^{(m)}$, that is, $D_{\mathrm{KL}}(\boldsymbol{\alpha}_i^{(d)} \| \boldsymbol{\alpha}_i^{(m)})$
- Outcome 4: KL divergence of $\boldsymbol{\alpha}_i^{(m)}$ from $\boldsymbol{\alpha}_i^{(d)}$, that is, $D_{\mathrm{KL}}(\boldsymbol{\alpha}_i^{(m)} \| \boldsymbol{\alpha}_i^{(d)})$
- Outcome 5: Disagreement ratio, which is the fraction of ads that have nonzero share under only one of the two policies in the set of all feasible ads.

The first four measures capture the extent of difference between ad shares, whereas the fifth measure uses a more binary approach and compares distributions in the set of ads that could be shown. We use these measures of discrepancy between the two policies and regress them on the set of historical features used in the previous section. Like before, we account for user and hour fixed effects. We present our results in Table 6, where each column shows how historical features are associated with each of the discrepancy measures. First, we find a consistently positive coefficient for the *number of past impressions*, which indicates that a richer history is associated greater differentiation between policies. Second, when we focus on the *variety of ads seen*, we find some weak negative links for the first four measures, and a strong negative link for the fifth measure. This is likely because higher variety of prior ads reduces the effective size of the action space by identifying poor-performing ads.

Third, we find that the *number of past clicks* is associated with more similar shares between the two policies (negative and significant coefficients in columns (1)–(4)), but not associated with any difference in the

Table 6. Discrepancy in the Distribution of Ad Allocation Between Dynamic Across the Historical Features

	Dependent variable: *Discrepancy between Ad Distributions under Dynamic and Myopic*				
	(1)	(2)	(3)	(4)	(5)
Number of past impressions	0.00043***	0.00019***	0.00111***	0.00053***	0.00005***
	(45.43)	(48.44)	(46.27)	(59.47)	(23.16)
Variety of ads seen	−0.00099*	0.00022	0.00109	−0.00096**	−0.00180***
	(−2.57)	(1.41)	(1.10)	(−2.62)	(−19.56)
Number of past clicks	−0.01496***	−0.00734***	−0.02471***	−0.01568***	0.00028
	(−19.22)	(−22.85)	(−12.39)	(−21.38)	(1.52)
Time since last session	−0.00008	−0.00005*	0.00006	−0.00009	−0.00002
	(−1.32)	(−2.10)	(0.38)	(−1.57)	(−1.03)
Last session length	−0.00095***	−0.00041***	−0.00068***	−0.00050***	0.00020***
	(−15.61)	(−16.14)	(−4.33)	(−8.65)	(13.56)
User fixed effects	✓	✓	✓	✓	✓
Hour fixed effects	✓	✓	✓	✓	✓
No. of observations	190,206	190,206	190,206	190,206	190,206
R^2	0.195	0.196	0.162	0.203	0.192
Adjusted R^2	0.138	0.139	0.103	0.147	0.135

Note. Numbers in parentheses are t statistics that are estimated using OLS.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

effective set of ads with a nonzero probability (insignificant coefficient in column (5)). This is likely because the existence of past clicks does not necessarily change the effective action space, but substantially increases the probability of a particular set of ads across both policies (e.g., ads similar to the ad that is already clicked on). Fourth, coefficients for our measure of recency (*time since the last session*) are all insignificant, indicating no association between usage recency and discrepancy between policies.

Fifth, we examine the link between the last session length and the discrepancy measures. In general, we expect a longer session to increase the discrepancy between the two policies because the *fully dynamic* policy has richer dynamics and more opportunities to differentiate. Surprisingly, we find that a higher session length is associated with more similar shares (columns (1)–(4)) but more disagreement in the set of ads that could be shown (column (5)). One potential explanation is that the discrepancy captured by our first four measures is more pronounced if the session is short. That is, although a longer session makes the set of ads more different, the probabilities become closer as they capture the specifics of leave probabilities.

In Online Appendix J, we examine what constitutes the discrepancy between the two policies in terms of their use of two session-level features that are widely used in the advertising literature: frequency and spacing.

5.4. Robustness Checks

We run a series of tests to check the robustness of the results presented in previous sections. We first establish the robustness of our results to different initializations of our framework, such as the number of ads in the action space (Online Appendix K.1), length of horizon (Online Appendix K.2), and the specific modeling and evaluation data sets used (Online Appendix K.3). We replicate our main qualitative results with these different initializations. We further demonstrate the robustness of our results by comparing the performance of our framework to other benchmarks, such as the one presented in Sun et al. (2017) and a predefined sequencing policy that does not use real-time information in Online Appendix K.4. Finally, we present the robustness of the U-shaped pattern in Figure 5 and our results in Table 5 to alternative specifications in Online Appendix K.5.

6. Implications

Our findings have several implications for managers and marketing practitioners, as we focus on the problem of value creation in advertising marketplaces. In particular, we demonstrate that incorporating within-session dynamics through our adaptive ad sequencing framework creates value in the marketplace by enhancing user engagement with ads compared with a series of benchmark policies such as the single ad policy that mimics the case for a nonrefreshable ad slot and adaptive myopic policy, which is the dominant allocation strategy used by firms (Theocharous et al. 2015). To that end, our findings have important applications for the publishers on what ad format to use (refreshable or nonrefreshable), and more importantly, what kind of allocation policy to adopt (myopic versus forward-looking). Specifically, our results suggest that the industry standard (adaptive myopic policy) leaves considerable value on the table, thereby calling for a change in the current practice in the industry, particularly because the computational cost of our framework is only slightly higher than an adaptive myopic framework.[20]

It is worth emphasizing that the framework is general and all ad platforms can use our framework to measure the gains in user engagement from adopting a fully dynamic framework, as long as there is unconfounded randomization in ad allocation. Although ad platforms often use deterministic auctions such as first- or second-price auctions for ad allocation, they can still incorporate some level of randomization through ϵ-greedy approaches (Theocharous et al. 2015) or small-scale experimentation (Ling et al. 2017). Similarly, platforms can use other measures of user engagement as the reward function and different optimization horizon, depending on the context. Thus, the applicability of our framework does not depend on the specific empirical setting in this paper.

Our sequencing framework can also be readily implemented in cases where a platform wants to sequence content to achieve optimal user-level outcomes. In particular, the improvement in ad response as a result of sequencing motivates a wide range of marketing applications that are closely related to advertising, such as sequencing promotional emails and notifications in an online retail context, sequencing articles in news websites to increase audience engagement, sequencing social media posts to enhance user experience, and sequencing push notifications for churn management. More broadly, our framework can be extended to other contexts where we want to use persuasive messaging through adaptive interventions. For example, in the context of mobile health, a growing body of work focuses on just-in-time adaptive interventions (JITAI) in mobile apps and studies their impact in shaping consumers' health behavior, including physical fitness and activity, smoking, alcohol use, and mental illness (Nahum-Shani et al. 2017). Similarly, in the context of education, these adaptive interventions can be used to improve students' motivation and outcomes (Mandel et al. 2014). These showcases can also inspire the public sector to use these tools in cases where collective action is required, such as environmental protection and political participation.

7. Conclusion

Mobile in-app advertising has grown exponentially over the last few years. The ability to exploit the

time-varying information about a user to personalize ad interventions over time is a key factor in the growth of in-app advertising. Despite the dynamic nature of the information, publishers often use myopic decision-making frameworks to select ads. In this paper, we examine whether a dynamic decision-making framework benefits the publisher in terms of the user engagement with ads, as measured by the number of clicks generated per session. Our dynamic framework has three main components: (1) a theoretical framework that models the domain structure such that it captures intertemporal tradeoffs in the ad allocation decision, (2) an empirical framework that breaks the policy identification problem into a combination of machine learning tasks that achieve sequence personalization, counterfactual validity, and scalability, and (3) a policy evaluation method that is separate from policy identification, thereby allowing a robust counterfactual policy evaluation. We apply our framework to large-scale data from the leading in-app ad network of an Asian country. Our results indicate that our adaptive ad sequencing policy results in significant gains in the expected number of clicks per session compared with a set of benchmark policies. In particular, we show that our policy results in 5.76% more clicks, on average, compared with the adaptive myopic policy that is the current state of practice. Almost all these gains stem from an increase in average response to each impression instead of increased usage by each user. Next, we document extensive heterogeneity in gains from adaptive ad sequencing and find a U-shaped pattern for gains over the length of users' past history, indicating that gains are highest for either new users or those whose past data are rich. As for the policy difference between adaptive ad sequencing and adaptive myopic, we find that our policy results in a greater ad diversity, which can be because our policy better manages user attention by showing a more diverse set of ads.

Our paper makes several contributions to the literature. First, from a methodological point-of-view, we develop a unified dynamic framework that starts with a theoretical framework that specifies the domain structure in mobile in-app advertising and an empirical framework that breaks the problem into tasks that can be solved using a combination of machine learning methods and causal inference tools. Notably, the BIQFA algorithm in our framework achieves scalability without imposing simplifying assumptions on the dynamics of the problem. Second, from a substantive standpoint, we document the gains from adopting an adaptive forward-looking sequencing policy. In particular, we show a 5.76% gain in clicks from adopting our fully dynamic policy over the adaptive myopic policy, and establish its robustness across a series of robustness checks. This comparison is of particular importance as the adaptive myopic policy is currently the standard approach in the industry. We further present a comprehensive study of heterogeneity and document key differences between our policy and adaptive myopic policy, which is of great value to managers who need to interpret the gains and understand when and why the framework is most valuable.

Nevertheless, our study has some limitations that serve as excellent avenues for future research. First, our counterfactual policy evaluation is predicated on the assumption that users do not change their behavior in response to sequencing policies. Although we exploit randomization to obtain our counterfactual estimate, it would be important to validate these findings in a field experiment. Furthermore, we use the training data offline to learn counterfactual estimates for click and leave outcomes. Extending our framework to an online setting that captures exploration/exploitation tradeoffs is important because online approaches are more cost-efficient and robust to cold-start problems. Finally, we use the entire within-session history to update state variables. Future research can look into more parsimonious frameworks that can be scalable to longer time horizons.

Acknowledgments

The author thanks committee chair and advisor Hema Yoganarasimhan and committee members Arvind Krishnamurthy, Simha Mummalaneni, Amin Sayedi, and Jacques Lawarree for guidance and comments; an anonymous firm for providing the data; the UW-Foster High Performance Computing Laboratory for providing computing resources; the selection committee for MSI Alden G. Clayton Doctoral Dissertation Award, Vithala R. and Saroj V. Rao ISMS Doctoral Dissertation Award, and American Statistical Association Doctoral Research Award (Statistics in Marketing Section) who supported the dissertation that this paper originated from; and the participants of the research seminars at University of Wisconsin–Madison, University of Colorado–Boulder, University of Southern California, University of Texas at Dallas, Texas A&M University, Harvard Business School, Stanford University, Yale University, University of Toronto, Penn State University, University of Rochester, Johns Hopkins University, Rutgers University, Carnegie Mellon University, Cornell Tech, Cornell University, University of California–San Diego, and Dartmouth College for feedback.

Endnotes

[1] A session is an uninterrupted time that a user spends inside an app. This is in contrast with the common practice in desktop advertising, where ads remain fixed throughout a session.

[2] In this paper, we use the publisher, ad network, and platform interchangeably when we refer to the agent who makes the ad placement decision.

[3] For an excellent summary of the current practice in ad allocation at major platforms, please see Despotakis et al. (2021). The auctions in place generally alternate between second-price, first-price, and VCG. None of these auctions uses a forward-looking allocation rule.

[4] Please see Rafieian and Yoganarasimhan (2022a) for an extensive summary of the literature on personalization.

[5] Please see chapter 7 in Tellis (2003) for a summary of the earlier work on advertising dynamics.

[6] There are obviously various ways to define a session based on the time gap between two consecutive exposures. We show that our results are robust to different definitions.

[7] We do not have the data on the banner creatives and its format, that is, whether it is a jpeg file or an animated gif.

[8] Our sampling procedure is almost identical to that of Rafieian and Yoganarasimhan (2022b). However, the number of impressions and sessions is slightly different because we need to drop users with missing information on latitude and longitude. Rafieian and Yoganarasimhan (2022b) use those impressions because latitude and longitude do not play a role in their analysis.

[9] In many contexts, the publisher can choose a no-ad option, where the impression is not filled with an ad. Because all ad opportunities are filled in our setting, we exclude the no-ad option from our action set. However, future research could easily extend our framework to include the no-ad option, depending on the empirical context.

[10] For a deterministic policy, $\pi(a \mid s)$ will take value one only for one ad for any given state.

[11] Naturally, we cannot use any information from the future to generate a feature: at any point, we only use the prior history up to that point.

[12] It is worth noting that the subscript t in $\tilde{Q}_t$ is only for notational simplicity.

[13] Formally, we can incorporate that by setting $\tilde{Q}_s(\cdot) = 0$ for any $s > T$.

[14] Because this is a supervised learning task, a huge discrepancy between the sample of states used for function approximation and the sample under the optimal policy can affect the performance of the q-function learned.

[15] We need to stress that although initialization helps the algorithm achieve higher efficiency, the algorithm works under alternative initialization approaches. For example, we yield the same result with an initialization of states under a fully random policy, but we need to use roughly double the size of sampled states for each t.

[16] Each one of these top 15 ads has been shown at least in 1% of all impressions.

[17] The closest approach to ours is Wilbur et al. (2013), who use more contextual and behavioral information to estimate continuation probabilities. We extend that approach by using a richer set of features and a more flexible learner.

[18] We further formalize these benchmark policies in Online Appendix I.1 and discuss the time complexity of identifying each policy and the *Fully Dynamic* policy in Online Appendix I.2.

[19] Each quintile contains 20% of all sessions, with quintile 1 being the bottom 20% of values.

[20] It is worth noting that our framework is readily applicable to non-strategic environments where the publisher wants to maximize user engagement, such as allocating impressions in contexts where ads are sold in bulk in prenegotiated reservation contracts. In real-time bidding auction environments where advertisers can strategically respond to the change in allocation, we need to design strategy-proof auctions that achieve the publisher's objective. Rafieian (2020) studies these strategic environments and proposes a revenue-optimal auction for adaptive ad sequencing.

References

Aguirregabiria V, Mira P (2002) Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica* 70(4):1519–1543.

Ansari A, Mela CF (2003) E-customization. *J. Marketing Res.* 40(2): 131–145.

Aravindakshan A, Naik PA (2011) How does awareness evolve when advertising stops? The role of memory. *Marketing Lett.* 22(3):315–326.

Arnosti N, Beck M, Milgrom P (2016) Adverse selection and auction design for Internet display advertising. *Amer. Econom. Rev.* 106(10):2852–2866.

Bellman R (1966) Dynamic programming. *Science* 153(3731):34–37.

Bellman R, Dreyfus S (1959) Functional approximations and dynamic programming. *Math. Tables Other Aids Comput.* 13(68):247–251.

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. Krishnapuram B, ed. *Proc. 22nd ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 785–794.

Despotakis S, Ravi R, Sayedi A (2021) First-price auctions in online display advertising. *J. Marketing Res.* 58(5):888–907.

Dubé J-P, Hitsch GJ, Manchanda P (2005) An empirical model of advertising dynamics. *Quant. Marketing Econom.* 3(2):107–144.

eMarketer (2018) Mobile in-app ad spending. Accessed April 19, 2018, https://forecasts-na1.emarketer.com/584b26021403070290f93a5c/5851918a0626310a2c 186a5e.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29(5):1189–1232.

Fu J, Kumar A, Soh M, Levine S (2019) Diagnosing bottlenecks in deep q-learning algorithms. Chaudhuri K, Salakhutdinov R, eds. *Proc. 36th Internat. Conf. Machine Learn.*, vol. 97, Proceedings of Machine Learning Research Series (PMLR), 2021–2030.

Goli A, Reiley D, Zhang H (2021) Personalized versioning: Product strategies constructed from experiments on pandora. Preprint, submitted July 8, last revised September 29, https://dx.doi.org/10.2139/ssrn.3874243.

Gordon GJ (1995) Stable function approximation in dynamic programming. *Machine Learning Proc.* (Elsevier, New York), 261–268.

Han S, Jung J, Wetherall D (2012) A study of third-party tracking by mobile apps in the wild. Technical report UW-CSE-12-03-01, University of Washington, Seattle.

Hasselt H (2010) Double q-learning. Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, eds. *Adv. Neural Inform. Processing Systems*, vol. 23 (Curran Associates, Inc., Red Hook, NY), 2613–2621.

Horsky D (1977) An empirical analysis of the optimal advertising policy. *Management Sci.* 23(10):1037–1049.

IAB (2021) 2020/2021 IAB internet advertising revenue report. Accessed April 7, 2021, https://www.iab.com/insights/internet-advertising-revenue-report/.

Jeziorski P, Segal I (2015) What makes them click: Empirical analysis of consumer demand for search advertising. *Amer. Econom. J. Microeconom.* 7(3):24–53.

Kallus N, Uehara M (2020) Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Machine Learn. Res.* 21:167–1.

Kar W, Swaminathan V, Albuquerque P (2015) Selection and ordering of linear online video ads. Werthner H, Zanker M, conference chairs. *Proc. 9th ACM Conf. on Recommender Systems* (ACM, New York), 203–210.

Kempe D, Mahdian M (2008) A cascade model for externalities in sponsored search. Papadimitriou C, Zhang S, eds. *Proc. Internat. Workshop on Internet and Network Econom.* (Springer, Berlin), 585–596.

Kristianto D (2021) Winning the attention war: Consumers in nine major markets now spend more than four hours a day in apps. Accessed April 8, 2021, https://www.appannie.com/en/insights/market-data/q1-2021-market-index/.

Le H, Voloshin C, Yue Y (2019) Batch policy learning under constraints. Chaudhuri K, Salakhutdinov R, eds. *Proc. 36th Internat. Conf. on Machine Learn.* (PMLR), 3703–3712.

Lee K, Laskin M, Srinivas A, Abbeel P (2021) Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning.

Meila M, Zhang T, eds. *Proc. 38th Internat. Conf. on Machine Learn.* (PMLR), 6131–6141.

Levine S, Kumar A, Tucker G, Fu J (2020) Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Preprint, submitted May 4, https://arxiv.org/abs/2005.01643.

Ling X, Deng W, Gu C, Zhou H, Li C, Sun F (2017) Model ensemble for click prediction in Bing search ads. Barrett R, Cummings R, chairs. *Proc. 26th Internat. Conf. on World Wide Web Companion* (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE), 689–698.

Little JD (1979) Aggregate advertising models: The state of the art. *Oper. Res.* 27(4):629–667.

Lu S, Yang S (2017) Investigating the spillover effect of keyword market entry in sponsored search advertising. *Marketing Sci.* 36(6):976–998.

Manchanda P, Dubé J-P, Goh KY, Chintagunta PK (2006) The effect of banner advertising on Internet purchasing. *J. Marketing Res.* 43(1):98–108.

Mandel T, Liu Y-E, Levine S, Brunskill E, Popovic Z (2014) Offline policy evaluation across representations with applications to educational games. Bazzan A, Huhns M, chairs. *Proc. Internat. Conf. on Autonomous Agents and Multi-Agent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC), 1077–1084.

Mannor S, Simester D, Sun P, Tsitsiklis JN (2007) Bias and variance approximation in value function estimates. *Management Sci.* 53(2):308–322.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, et al. (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Mullainathan S, Spiess J (2017) Machine learning: An applied econometric approach. *J. Econom. Perspective* 31(2):87–106.

Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA (2017) Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann. Behav. Medicine* 52(6): 446–462.

Naik PA, Mantrala MK, Sawyer AG (1998) Planning media schedules in the presence of dynamic advertising quality. *Marketing Sci.* 17(3):214–235.

Nerlove M, Arrow KJ (1962) Optimal advertising policy under dynamic conditions. *Economica* 29(114):129–142.

Rafieian O (2020) Revenue-optimal dynamic auctions for adaptive ad sequencing. Working paper, Cornell Tech, New York.

Rafieian O, Yoganarasimhan H (2021) Targeting and privacy in mobile advertising. *Marketing Sci.* 40(2):193–218.

Rafieian O, Yoganarasimhan H (2022a) AI and personalization. Preprint, submitted June 10, https://dx.doi.org/10.2139/ssrn.4123356.

Rafieian O, Yoganarasimhan H (2022b) Variety effects in mobile advertising. *J. Marketing Res.* 59(4):718–738.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

Rutz OJ, Bucklin RE (2011) From generic to branded: A model of spillover in paid search advertising. *J. Marketing Res.* 48(1):87–102.

Sahni NS (2015) Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quant. Marketing Econom.* 13(3):203–247.

Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J. Res. Development* 3(3):210–229.

Sawyer AG, Ward S (1979) Carry-over effects in advertising communication. *Res. Marketing* 2:259–314.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Sci.* 36(4):500–522.

Simester DI, Sun P, Tsitsiklis JN (2006) Dynamic catalog mailing policies. *Management Sci.* 52(5):683–696.

Simon H (1982) ADPULS: An advertising model with wearout and pulsation. *J. Marketing Res.* 19(3):352–363.

Sun Z, Dawande M, Janakiraman G, Mookerjee V (2017) Not just a fad: Optimal sequencing in mobile in-app advertising. *Inform. Systems Res.* 28(3):511–528.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Tellis GJ (2003) *Effective Advertising: Understanding When, How, and Why Advertising Works* (Sage Publications, Thousand Oaks, CA).

Theocharous G, Thomas PS, Ghavamzadeh M (2015) Personalized ad recommendation systems for life-time value optimization with guarantees. Yang Q, Wooldridge M, eds. *Proc. 24th Internat. Joint Conf. on Artificial Intelligence* (AAAI Press, Menlo Park, CA), 1806–1812.

Thomas P, Brunskill E (2016) Data-efficient off-policy policy evaluation for reinforcement learning. Balcan MF, Weinberger KQ, eds. *Proc. Internat. Conf. on Machine Learn.* (PMLR), 2139–2148.

Thomas P, Theocharous G, Ghavamzadeh M (2015) High-confidence off-policy evaluation. *Proc. AAAI Conf. on Artificial Intelligence*, vol. 29 (AAAI Press, Menlo Park, CA), 3000–3006.

Thomas PS, da Silva BC, Barto AG, Giguere S, Brun Y, Brunskill E (2019) Preventing undesirable behavior of intelligent machines. *Science* 366(6468):999–1004.

Tsitsiklis JN, Van Roy B (1996) Feature-based methods for large scale dynamic programming. *Machine Learn.* 22(1):59–94.

Urban GL, Liberali G, MacDonald E, Bordley R, Hauser JR (2013) Morphing banner advertising. *Marketing Sci.* 33(1):27–46.

Van Hasselt H, Doron Y, Strub F, Hessel M, Sonnerat N, Modayil J (2018) Deep reinforcement learning and the deadly triad. Preprint, submitted December 6, https://arxiv.org/abs/1812.02648.

Wilbur KC (2008) A two-sided, empirical model of television advertising and viewing markets. *Marketing Sci.* 27(3):356–378.

Wilbur KC, Xu L, Kempe D (2013) Correcting audience externalities in television advertising. *Marketing Sci.* 32(6):892–912.

Yi J, Chen Y, Li J, Sett S, Yan TW (2013) Predictive model performance: Offline and online evaluations. Ghani R, Senator TE, Bradley P, Parekh R, He J, eds. *Proc. 19th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 1294–1302.

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Sci.* 66(3):1045–1070.

Yoganarasimhan H, Barzegary E, Pani A (2022) Design and evaluation of personalized free trials. *Management Sci.*, ePub ahead of print August 10, https://doi.org/10.1287/mnsc.2022.4507.

Zantedeschi D, Feit EM, Bradlow ET (2017) Measuring multichannel advertising response. *Management Sci.* 63(8):2706–2728.