DISPLAY AD MEASUREMENT USING OBSERVATIONAL DATA: A REINFORCEMENT LEARNING APPROACH

By

Srinivas Tunuguntla

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Business)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 05/10/2022

The dissertation is approved by the following members of the Final Oral Committee:

Neeraj Arora, Professor, Business
Paul R. Hoban, Economist, Amazon.com
Qing Liu, Associate Professor, Business
Kevin Chung, Associate Professor, Business
Robert D. Nowak, Professor, Engineering
Alan Sorensen, Professor, Economics

TABLE OF CONTENTS

ABSTRACT

This paper introduces an observational framework for measuring the effects of display advertising. Causal measurement of display ad effects using observational data presents a significant challenge due to the dynamic nature of the data generating process. The proposed framework, derived from Markov Decision Processes, accurately reflects the data generating process and accounts for the different sources of endogeneity by explicitly modeling the users' browsing behavior and the advertiser's decision making. Using the proposed framework, we develop a novel estimation method that recovers the incremental impact on outcomes attributable to a display advertising campaign. We validate the proposed method using a randomized controlled trial.

Keywords: display advertising; ad effect measurement; Markov decision processes; reinforcement learning

1. INTRODUCTION

Display advertising budgets have grown rapidly over the past several years. The advent of real-time bidding (RTB) has enabled advertisers to target individual users and control the timing and location of each exposure. Advertisers view such increased precision as the key driver of return on investment (IAB 2021).

The rapid growth of display advertising has resulted in the need to accurately estimate its impact. Assessing the incremental impact of impressions on outcomes allows advertisers to make informed decisions on budget allocation for the display channel, and the various campaign and targeting strategies therein. To measure the effects of display advertising, advertisers must estimate the counterfactual outcomes had the users not seen any ads.

Despite advances in targeting, such causal measurement remains a challenging issue. While randomized controlled trials (RCTs) are the "gold standard", firms are often unable or unwilling to implement them due to the large costs associated with experimentation (Gordon et al. 2021). Instead, firms rely on observational data collected during the normal course of a campaign.

Observational methods, however, face challenges due to the dynamic nature of the data generating process. Unlike the traditional causal model (Rubin 2005), for each user in RTB, advertisers have multiple opportunities to serve an impression. At each impression opportunity, advertisers use information on the user's browsing, exposure, and purchase history to compute their willingness to serve an impression. As a result, a user's probability of exposure evolves dynamically throughout the campaign. In addition to the exposures, a user's outcomes—conversion metrics such as purchases and website visits—are also dynamic. Because each user can be exposed to multiple impressions, the timing and the temporal spacing between them jointly determine the user's outcomes (Sahni 2015). To measure the causal impact of a campaign,

observational approaches must take these dynamics into account while modeling the data generating process. Indeed, Gordon et al. (2019) and Lewis et al. (2011) find that, when a static model of assignment and outcomes is assumed, observational approaches can yield significantly biased estimates.

In this paper, we propose a novel observational approach to measure the incremental impact of an advertising campaign. We model the data generating process using a Markov Decision Process (MDP). In the model, each user is described by a *state* that evolves over time as a Markov process. At each time step, the user state—which follows the Markov property—captures all the information required to describe the user's current and future behavior. The advertiser observes partial information about this state—for example, through browsing and exposure history—and drives conversions by inducing alterations to the state through impressions.

The MDP framework accurately reflects the dynamic nature of the data generating process. At each time period, it provides a way to explicitly model (1) a user's probability of exposure—as a policy that maps the information available on the user to the advertiser's intent to serve an impression, (2) the users' response to impressions—via the state transition probabilities, and (3) outcomes—as a mapping from the user state to the probability of conversion.

The identification strategy relies on an intrinsic property of the MDP that the state transition probabilities are invariant to the advertiser's policy (Rust 1994, Hotz and Miller 1993). In other words, the state transition probabilities are structural parameters with respect to the advertiser's policy (Reiss and Wolak 2007). Thus, outcomes of counterfactual policies are identified through these transition probabilities. In Sections 3 and 4, we show that the causal impact of a campaign can be estimated by computing outcomes of one such counterfactual policy. We estimate these counterfactual outcomes by leveraging techniques from reinforcement learning.

We validate the estimate provided by the proposed method through a large-scale field experiment where users are randomly assigned to either a treatment or a control group. The users in the treatment group are eligible for exposure and are served impressions using the advertiser's targeting policy. The users in the control group are not eligible to receive impressions. We show that the proposed method recovers the counterfactual outcomes, using only data from the treatment group. Specifically, we show that the proposed method's estimate of the percentage lift in conversions attributable to the campaign is within 6% of the estimate obtained through the RCT. Moreover, the difference between the two estimates is about .31 times the standard error of the RCT estimate. In contrast, we show that the methods that do not account for the dynamics of the context yield significantly biased estimates, consistent with the literature.

We contribute to the display advertising literature by providing a novel approach to ad effect measurement. This approach uses readily available observational data, avoiding the large costs associated with RCTs. The approach also generalizes to any context involving dynamic treatments and is capable of computing outcomes of counterfactual treatment policies. Thus, we also contribute to the growing literature at the intersection of causal inference and machine learning by showing that techniques from reinforcement learning can be applied for inference in such settings.

The rest of the document is organized as follows. We introduce the real time bidding ecosystem and review the relevant literature on causal methods and MDPs in Section 2. In Sections 3 and 4, we present the MDP framework and an estimation method to recover counterfactual outcomes respectively. Section 5 describes the field experiment and validates the estimate provided by the proposed method. Section 6 concludes.

2. BACKGROUND AND RELATED LITERATURE

2.1. The Real Time Bidding Ecosystem

The RTB ecosystem is a two-sided marketplace with advertisers and publishers. Publishers manage webpages with content in which users are primarily interested. They supply ad inventory on these webpages to advertisers to generate revenue. Advertisers purchase these ad impressions to promote their brand, product, or service. There are many intermediaries, such as ad exchanges, demand and supply side platforms, that provide fundamental infrastructure to facilitate selling, buying, and serving ads in real time. For the sake of clarity, we limit our discussion to advertisers, publishers, and ad exchanges.

When a user visits a webpage on a publisher's website, a bid request is triggered for an impression opportunity. The publisher makes its impression available through an ad exchange which then queries all participating advertisers for bids on the impression. Advertisers evaluate the impression opportunity using information available from various sources. This includes contextual information (e.g., domain name of the publisher, topic, keywords), behavioral information such as browsing history tracked using cookies, and demographic information obtained through third parties. Each advertiser then submits a sealed bid to the ad exchange based on their private valuations. The ad exchange runs an auction and determines the winner. The winning advertiser's ad is then displayed to the user on the publisher's webpage.

From the advertiser's perspective, a campaign is a series of auctions corresponding to users that satisfy pre-determined matching criteria such as geographic location, device type, and language. A typical campaign duration can range from several weeks to a few months. Advertisers face challenges due to the volume and velocity of the auctions in the RTB ecosystem. Ad exchanges process, on average, 1.6 million auctions per second (Shen et al. 2015) and require

advertisers to submit bids within milliseconds for each auction (Google 2021b). As a result, advertisers manage their campaigns via automated targeting and bidding algorithms (Tunuguntla and Hoban 2021).

2.2. Causal Measurement

The objective of the advertiser is to drive conversions—user driven actions such as clicks, pageviews, and purchases. The causal impact of a campaign is therefore measured through its incremental impact on the total number of conversions. To measure this, advertisers must compare the conversions from a campaign and the counterfactual conversions that would have transpired had they not run a campaign. Researchers have proposed both experimental and observational approaches to measure the causal impact of display advertising.

2.2.1. Experimental Approaches

Experimental approaches or randomized controlled trials (RCTs) randomly assign users to treatment and control groups and compare the conversions between these groups. The users in the control group are not eligible for exposure. Although all the users in the treatment group are eligible, not all of them receive an impression. The advertiser's targeting algorithm determines the bids submitted to each auction depending on the user's browsing/purchase history, timing, and location of the impression. Thus, only a subset of the users in the treatment group are exposed.

Earlier approaches (Lewis and Rao 2015, Hoban and Bucklin 2015) delivered a public service announcement (PSA) to the control group users instead of the focal advertiser's impression. These approaches, however, are expensive due to the cost of PSAs and require coordination among advertisers and third-party charities. Moreover, when advertisers use computer algorithms to optimize ad delivery separately for the PSA and focal campaigns, these approaches produce biased estimates (Johnson et al. 2017).

Recent approaches (Gordon et al. 2019) reduce the cost of PSAs by withdrawing from RTB auctions corresponding to the control group users. As a result, the control group users are never exposed to impressions from the focal advertiser. Johnson et al. (2017) show that one of the limitations of these approaches is that they provide imprecise estimates because the comparison of conversions from the treatment and control groups includes unexposed users. They propose a method involving "ghost ads" to improve measurement precision by identifying users in the control group that would have been exposed had they been in the treatment group. The ghost ads framework, however, requires coordination among advertisers, demand side platforms, and ad exchanges.

Despite recent advances in reducing the cost, experimental approaches remain expensive due to the opportunity cost of not reaching the control group. The advertiser foregoes revenue by not serving impressions to these users who are otherwise as attractive to the firm as the treatment group users. This opportunity cost can be large because of the sample size needed for precise measurement. Lewis and Rao (2015) show that informative experiments can require tens of millions of observations due to low statistical power. Moreover, advertisers typically manage a large number of campaigns varying on products advertised, target markets, conversion events, and budgets allocated. Such diverse campaigns exhibit a large heterogeneity in their incremental impacts, requiring advertisers to practice continuous experimentation (Zantedeschi et al. 2017). Thus, the opportunity cost of implementing experiments is recurrent and is associated with each campaign.

In addition to the large sample size requirements, experimental approaches suffer from a more serious limitation for budget constrained advertisers: they produce biased estimates of the potential return on investment. We illustrate this using the following stylized example. Consider a

scenario where there are three users—1, 2, 3 with valuations v_1, v_2, v_3—in the treatment group and three users—1', 2', 3' with identical valuations v_1, v_2, v_3 —in the control group. The valuations denote the incremental lift in the probability of conversion resulting from an impression for each user. Let $v_1 > v_2 > v_3$. With an advertiser who has a budget to serve only two impressions, users 1 and 2 are exposed. The measured ROI is thus equal to $v_1 + v_2$. In the absence of the control group, however, users 1 and 1'—who have the highest valuations—are exposed. The maximum potential ROI of the campaign is then $2v_1$. Thus, although RCTs provide an unbiased estimate of the impact on exposed users from the treatment group, they understate the potential impact of advertising. That is, they induce a trade-off between measurement and maximizing campaign impact. Given the prevalence of budgets (Choi et al. 2020) and that the primary motivation of causal measurement is to compute the return on investment and to inform budget allocation, we consider this limitation a major impediment to the deployment of RCTs in the context of display advertising.

Due to these challenges, most firms either do not or cannot measure ad effects using experimental or quasi-experimental methods (Gordon et al. 2021). Instead, they rely on readily available observational data. Observational methods can be implemented with no additional cost or requirements on coordination with third party entities. Further, observational approaches do not suffer the same trade-off between measuring and maximizing campaign impact as experimental approaches.

2.2.2. Observational Approaches

Under an observational approach, all the users are eligible for exposure. Similar to the treatment group in the experimental approaches, exposure of a given user is determined by the advertiser's targeting algorithm. Observational approaches use the data from all the users—

exposed and unexposed—to compute the counterfactual conversions that would have transpired had the advertiser not run a campaign.

In the absence of randomized assignment, observational approaches must account for the data generating process to estimate causal effects. Failure to do so can severely bias the estimates (Rubin 2005, Heckman and Pinto 2003). In RTB, this involves modeling the assignment of impressions and the distribution of conversions conditional on exposures. Modeling these, however, is challenging due to their dynamic nature and endogeneity concerns.

Because an auction is triggered whenever a user visits a publisher's website in RTB, advertisers have multiple opportunities to serve an impression for each user. At each auction, advertisers evaluate the impression opportunity using dynamic information such as browsing history, prior conversions, and publisher's contextual information. As a result, a user's probability of exposure evolves dynamically throughout the campaign period. Modeling the assignment mechanism, therefore, requires modeling a user's probability of exposure at each auction. Ignoring such dynamics can lead to known issues such as activity bias (Lewis et al. 2011), which is a result of a user's probability of exposure being a function of browsing intensity.

The probability of exposure also varies across users depending on factors such as demographics, browsing, and purchasing behavior. These factors induce systematic differences between exposed and unexposed users, resulting in endogeneity. One such source of endogeneity is the targeting criteria (Lewis and Rao 2015). Because media buyers are generally rewarded for having shown their ads to users who later convert, targeting algorithms typically target users that are most likely to convert (Choi et al. 2020). That is, exposed users are specifically chosen based on their higher conversion rates. Similarly, competitive effects through the auctions are another potential source of endogeneity. An exposure is determined not only by the focal advertiser's

valuation of the impression but also by other advertisers' valuations. The focal advertiser is likely to win impression opportunities that they value highly (Gordon et al. 2019). Observational approaches, therefore, must account for these differences between exposed and unexposed users to accurately measure advertising effectiveness.

Conversions in RTB are also dynamic. Because each user can be exposed to multiple impressions, their volume, timing, and the temporal spacing between them jointly determine the user's outcomes. Repeated ad exposures have wear-in and wear-out effects on the user and affect the marginal effectiveness of each impression (Braun and Moe 2013). Similarly, impressions have carryover effects, which are determined by temporal spacing between them (Sahni 2015). Thus, the probability of conversion during any time period depends on the schedule of past ad exposures. Accounting for this is important to measure the effect of advertising precisely (Sahni 2015).

Observational methods face challenges in measuring campaign effectiveness because of these distinctive features of RTB's data generating process. Gordon et al. (2019) and Lewis et al. (2011) explore the performance of observational methods commonly used by researchers and practitioners. They find that, when a static model of assignment is assumed and when the sources of endogeneity are not sufficiently controlled for, there is a large discrepancy between the estimates provided by observational and experimental approaches. Gordon et al. (2019) show that the estimates from observational approaches are off by at least a factor of three in more than half of the campaigns they analyzed. They also find that these methods mostly overestimate the impact of advertising, although in some cases, they significantly underestimate it.

In this work, we propose an observational method based on a Markov Decision Process (MDP) that reflects the data generating process. We model the assignment mechanism through the advertiser's decision making. At each auction, we compute the advertiser's willingness to serve an

impression—and therefore the probability of exposure—as a function of the history of observables. This accounts for the dynamic nature of impressions and controls for targeting criteria. Competition induced endogeneity can occur because a user's exposure depends not only on the focal advertiser's bid but also on other advertisers' bids. We eliminate such endogeneity by quantifying the impact of bids on conversions rather than the impact of impressions. We account for the temporal and carryover effects of impressions by estimating a "value function" that quantifies their effect on future conversions. Finally, we use flexible functional forms via neural networks to model all relationships, thus avoiding any bias resulting from model misspecification.

2.3. Markov Decision Processes

Markov Decision Processes or MDPs (Sutton and Barto 2018) provide a mathematical framework for modeling sequential decision making where a decision-making *agent* interacts sequentially with a dynamic *environment* in discrete time steps. The agent-environment interaction is shown in Figure 1. At each time step t, the environment is described by a state S_t. The state S_t contains all the relevant information to completely characterize the future trajectory of the environment. In other words, the environment state follows the Markov property: the distribution of future states is independent of the past states, conditional on the current state.

The agent receives a potentially noisy representation of this state, referred to as an observation O_t. Settings where the state S_t is not completely observable by the agent, i.e., the observation O_t is not equal to S_t, are referred to as partially observable Markov decision processes or POMDPs (Arulkumaran et al. 2017). In a POMDP, the observation O_t is characterized by a distribution $\mathcal{O}(O_t \mid S_t)$ that is conditional on the current state S_t.

The agent chooses an action A_t at each time step. In a POMDP, this action typically depends on the entire history of observations $H_t = \{O_1, O_2, \dots, O_t\}$. In part as a consequence of the

agent's action, the environment moves to a new state S_{t+1}, and the agent receives a numerical reward R_{t+1}. The state transition and the agent's reward are characterized by the joint conditional distribution $\mathcal{T}(S_{t+1}, R_{t+1} \mid S_t, A_t)$.

The agent's payoff is equal to the accumulated reward. Several measures of accumulated reward such as total cumulative reward, total discounted reward, and average long-term reward have been studied in the literature (Puterman 2014). The agent chooses actions such that the environment moves to favorable states that generate higher accumulated reward. The agent can control the environment's state transition because the transition dynamics $\mathcal{T}$ partially depend on the agent's action. The agent's decision making is characterized by a policy π that maps the history of observations to actions.

The MDP framework provides a generalization of goal directed behavior. Sutton and Barto (2018) state that any problem of goal directed behavior can be modeled using the three elements of an MDP: the environment's dynamics described by a Markov process, the agent's decision making characterized by a policy and the agent's objective. Indeed, MDPs have been successfully applied to many different contexts: solving games like Go (Silver et al. 2017) and poker (Bowling et al. 2017), learning control policies in robotics (Levine et al. 2016, 2018), option pricing (Tsitsiklis and Van Roy 2001), portfolio optimization (Moody and Saffell 2001, Deng et al. 2016), disease diagnosis (Peng et al. 2018), medical imaging (Li et al. 2018), personalized education (Upadhyay et al. 2018), and optimizing energy usage (Glavic et al. 2017).

MDPs are appropriately suited for modeling display ad campaigns. Psychological research has found that Markov models can be used to describe and formalize human behavior (Wickens 1982, Visser et al. 2002). The states of the Markov model are interpreted to be cognitive, emotional states that produce behavior, while the transition probabilities describe the evolution of these states

(Visser 2011). In particular, hidden Markov models have been found to accurately model browsing behavior (Awad and Khalil 2012, Scott and Hann 2006). Furthermore, Balachandran and Deshmukh (1976) and Hauser and Wisniewski (1982) show that persuasive communications such as impressions alter the user state. They also show that the effects of such persuasive communications can be modeled as an alteration to the user state transition probabilities. That is, the distribution of the new user state is conditionally dependent on exposure to an impression.

The MDP framework can therefore be applied to display ad campaigns by modeling the users as instantiations of the environment and the advertiser as the decision-making agent. The advertiser participates in the RTB auctions to drive conversions by inducing favorable transitions to the user state through impressions. In the MDP framework, the agent's actions are the advertiser's bids in the auctions and the rewards are conversions. The advertiser's payoff is the accumulated conversions.

3. MODEL

3.1. MDP Framework — Data Generating Process

During the course of a campaign, the advertiser interacts sequentially with each user i through RTB auctions indexed by n. We treat each user as an independent instantiation of the MDP. Accordingly, we drop any explicit references to an individual user i unless necessary.

The data generating process is summarized as a directed acyclic graph (DAG, Pearl 2000) in Figure 2. Each node in the graph represents a variable and the directed edges capture causal relationships between the variables. The latent variables in the model are denoted by dotted circles and the observable variables by solid circles.

The user state at auction n is denoted by S_n, which satisfies the Markov property. That is, the user state contains all the information required to describe the user's current and future behavior. We assume that the user state is latent and unobservable by the advertiser. The advertiser submits a bid based on the information available on the user. This includes contextual information of the auction (e.g., domain name of the publisher, topic, keywords), behavioral information such as browsing/purchase history tracked using cookies, and demographic information obtained through third parties. We refer to any information obtained by the advertiser at auction n as observation O_n. This observation is dependent on the user state S_n and is characterized by the conditional distribution $\mathcal{O}(O_n \mid S_n)$.

The advertiser's bid B_n typically depends on the entire history of observations denoted by $H_n = \{O_1, O_2, \dots, O_n\}$. The advertiser's decision making is characterized by a policy $\pi_A: \mathcal{H} \to \mathbb{R}$ that maps a history of observations to a bid. Thus, $B_n = \pi_A(H_n)$.

The objective of the advertiser is to drive conversions, which are user driven actions such as clicks, pageviews, and purchases. Let R_{n+1} be an indicator for user conversion during the period

between auctions n and $n+1$. That is, R_{n+1} is equal to 1 if the user converts between auctions n and $n+1$, and 0 otherwise. The joint distribution of R_{n+1} and the user's subsequent state S_{n+1} conditional on the current state S_n and the advertiser's bid B_n is given by $\mathcal{T}(S_{n+1}, R_{n+1} \mid S_n, B_n)$.

It is worth noting that we model the user state transitions conditional on the advertiser's bid, not an exposure. This eliminates any endogeneity induced by the auction mechanism. An exposure is determined not only by the focal advertiser's bid, but also by other participating advertisers' bids. These competing bids can depend on variables unobservable by the focal advertiser, which can induce selection into the focal ad exposures by systematically altering the set of auctions resulting in an exposure. However, because the focal advertiser's bid can only depend on variables observable by it, modeling the transition probabilities as a function of the bid eliminates any endogeneity induced by the auction mechanism.

The advertiser's payoff is equal to the accumulated conversions. While there are several measures of accumulated reward, we focus on the average conversions per auction as a measure of the advertiser's payoff. This allows us to normalize conversions across users, who might have different number of auctions associated with them due to differences in browsing intensity. We finally note that similar analysis can be derived for any other linear combination of the conversions at all periods (Puterman 2014).

We assume that the distributions characterizing the MDP—transition and conversion dynamics $\mathcal{T}$ and distribution of observations $\mathcal{O}$—are unknown. To compute counterfactual outcomes, we present an estimation method that is agnostic to these distributions and depends only on data observed from implementing the advertiser's policy π_A. This data constitutes the sequence of advertiser's observations $\{O_1, O_2, \dots\}$, bids $\{B_1, B_2, \dots\}$, and conversions $\{R_1, R_2, \dots\}$.

3.2. Identification

To estimate the causal impact of π_A, we must compute the counterfactual conversions in the absence of the campaign. We capture this counterfactual scenario by defining a null policy π_N . The null policy submits a constant bid of zero for each auction. Because this is equivalent to not running a campaign, we estimate the outcomes of implementing the null policy to determine the causal impact of the advertiser's policy.

As described above, the quality of a policy is measured through the average number of conversions obtained per auction. Each policy π is associated with a long-term average conversion rate ρ_π defined as

$$\rho_\pi = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[R_n \mid B_{0:n-1} \sim \pi] \tag{1}$$

where the expectations are conditioned on bids submitted according to policy π. The average conversion rate ρ_π depends on the steady state distribution of the Markov process while following π.

The incremental impact of a campaign can then be computed by comparing the average conversion rates resulting from the advertiser's policy π_A and the null policy π_N. Following the extant literature, we express the impact of a campaign as the percentage lift in the conversion rate.

$$lift(\pi_A) = \frac{\rho_{\pi_A} - \rho_{\pi_N}}{\rho_{\pi_N}} \times 100 \tag{2}$$

To estimate the lift, we must estimate both ρ_{π_A} and ρ_{π_N}. Because we observe data from the advertiser's policy, ρ_{π_A} can be estimated as the ratio of total number of conversions observed and the total number of auctions. We now show that ρ_{π_N} is identified through the observed data using

the framework of graphical causal models (Pearl 2000). We first note that ρ_{π_N} can be defined using do-calculus (Pearl 2000) as

$$\rho_{\pi_N} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[R_n \mid do(B_0 = 0, \dots, B_{n-1} = 0)] \tag{3}$$

where the notation $do(X = x)$ refers to intervening in the data generating process by artificially forcing the variable X to take value x and the variables inside the *do* operator are referred to as the intervention variables. In this case, ρ_{π_N} is the average conversion rate in the counterfactual scenario where all the advertiser bids are forced to be zero.

A sufficient condition for identification is that all the causal factors of the intervention variables are observed. When the data generating process is expressed as a DAG, this means that all the parents of the intervention variables are observed. This is known as the back-door criterion (Pearl 2000)—a generalization of the unconfoundedness assumption in Rubin's causal model. In RTB, the advertiser's bids are the intervention variables. Because bid decisions are made algorithmically—using information such as browsing/purchase history—it is common for advertisers to observe all the inputs to these bid decisions, satisfying the back-door criterion, by definition. In other words, it is sufficient to observe the history of observations H_n in Figure 2 for identification of ρ_{π_N} through the data observed from implementing the advertiser's policy π_A.

The estimation of ρ_{π_N}, however, also requires a sequential analogue of the overlap assumption, which we discuss in the following subsection.

3.3. Bellman Equation

Our estimation method relies on the fact that the data generating process satisfies a Bellman equation (Schwartz 1993). To show this, we first define a "value function"—henceforth referred to as the Q-function—that quantifies the effect of individual bids on the total number of

conversions. Consider an auction where the history of observations is h, a bid b was submitted. Assuming a policy π is implemented for all subsequent bids, the Q-function is defined as

$$Q_\pi(h, b) = \mathbb{E}_\pi\left[\sum_{k=1}^{\infty}(R_{n+k} - \rho_\pi) \mid H_n = h, B_n = b\right] \tag{4}$$

The Q-function—also known as the state-action value function in reinforcement learning literature (Sutton and Barto 2018)—measures the incremental impact of an individual bid on total conversions over the steady state average conversion rate while following policy π. The Q-function satisfies a Bellman equation (Schwartz 1993). The recursive relationship defined by the Bellman equation can be derived from Equation 4. The right-hand side of Equation 4 can be written as

$$\mathbb{E}[(R_{n+1} - \rho_\pi) \mid H_n = h, B_n = b] + \mathbb{E}_\pi\left[\sum_{k=2}^{\infty}(R_{n+k} - \rho_\pi) \mid H_n = h, B_n = b\right]$$

By definition, the second term is equal to $\mathbb{E}[Q_\pi(h', b')]$, where $H_{n+1} = h'$ is the history h appended with the observation $O_{n+1} = o$ and $b' = \pi(h')$. Thus, the Bellman equation is given by

$$Q_\pi(h, b) = \mathbb{E}[(r - \rho_\pi) + Q_\pi(h', b')] \tag{5}$$

where $R_{n+1} = r$ is the immediate reward received. The expectation is with respect to the transition dynamics of the MDP and the distribution of observations. Intuitively, the Bellman equation states that the incremental impact of a bid is equal to the immediate differential reward $(r - \rho_\pi)$ and the incremental impact of the subsequent bid.

The Bellman equation holds for any arbitrary policy π. The solution to the Bellman equation for $\pi = \pi_N$ provides us with an estimate of ρ_{π_N}. To gain some intuition on the estimation procedure, note that all bids are equal to zero under policy π_N. Therefore ρ_{π_n} encodes the impact

of zero bids on conversions. This information is provided by $Q_{\pi_N}(h, 0)$ and the relationship between the Q-function and the average conversion rate is established by the Bellman equation.

3.3.1. Overlap

It is clear that estimating $Q_{\pi_N}(h, 0)$ accurately for all h is necessary in order to estimate ρ_{π_N}. A sufficient condition to estimate $Q_{\pi_N}(h, 0)$ is that we observe $b = 0$ in the data for any h with non-zero probability. This is analogous to the overlap assumption in Rubin's causal model (Robins 2004). A key challenge in RTB, however, is that advertisers typically employ targeting algorithms that are deterministic. That is, there is typically no exogenous variation in bids given a history h.

We overcome this challenge by taking advantage of the exogenous variation induced by the auctions. Consider an auction where the focal advertiser lost with a bid b. The trajectory of the user thereafter would be the same had the advertiser submitted any bid less than b—including a bid of zero. We can use this observation to augment the data with a bid of zero whenever the focal advertiser loses an auction (the implementation details are discussed in Section 4.2). Thus, the overlap assumption is satisfied if there is a non-zero probability of the focal advertiser losing an auction given any history of observations h. This is a reasonable assumption in RTB, given the large number of advertisers with private information participating in any auction.

4. ESTIMATION

Our estimation procedure involves two stages. In the first stage, we reduce the dimensionality of the history of observations by computing predictive state representations (PSRs). These PSRs follow the Markov property and act as sufficient statistics for computing advertiser's bids and predicting a user's future trajectory. In the second stage, we implement an iterative algorithm that solves the Bellman equation to jointly estimate the Q-function and the average conversion rate of the counterfactual policy.

4.1. Predictive State Representation

Recall that the Q-function is computed at a history h for a bid b. While this can be estimated in theory, there is a potential computational challenge. Histories grow with the number of auctions per user and can become large and unwieldy. Therefore, for computational tractability, we first implement a dimensionality reduction step that maps each history to a finite dimensional representation. These finite dimensional representations act as a compact representative summary of the history.

To construct such a representative summary, we first note that any dynamical system can be completely characterized as a probability distribution over all possible future observations conditional on the past. Therefore, it is sufficient for the representative summary to follow the Markov property and be as predictive of the future observations as the actual history (Singh et al. 2012). Formally, let $Z_n = f(H_n)$ be the representative summary. The function f must be such that, for any histories h and h' that are mapped to the same representation—i.e., $f(h) = f(h')$—they also have the same probabilities for their next observation (Sutton and Barto 2018).

$$f(h) = f(h') \Longrightarrow \Pr(O_{n+1} = o \mid H_n = h, B_n = b) = \Pr(O_{n+1} = o \mid H_n = h', B_n = b)\ \forall o, b$$

Predictive state representations or PSRs (Littman et al. 2001, Thon and Jaeger 2015) provide a way to compute such representative summaries. They are constructed by minimizing a prediction loss between the predicted and the realized future observations. PSRs are widely used in modeling dynamical systems and construction of Markov state spaces (Downey et al. 2017a). Zhu et al. (2020) have recently applied them to marketing settings and found that the resulting summaries satisfy the Markov property and are even robust to non-Markov distortions of the data generating process.

Recurrent neural networks (RNNs) are aptly suited to compute a PSR (Downey et al. 2017b). RNNs model sequential data by maintaining an internal hidden state that is useful to predict future elements of the sequence. This internal state serves as the PSR. Moreover, RNNs provide a computationally convenient way to calculate the PSR using a recursive update u for every observation

$$Z_n = u(Z_{n-1}, B_n, O_n) \tag{6}$$

Note that, for any function f above that maps histories to PSRs, there always exists a corresponding recursive update u (Sutton and Barto 2018).

We build an RNN—referred to as the PSR network henceforth—as depicted in Figure 3. We provide the implementation details in Appendix A. The RNN has two modules, a recurrent and a predictive module. The recurrent module takes as input the tuple (O_n, B_n) and computes the PSR Z_n recursively, as shown in Equation 6. The predictive module predicts the subsequent observation $\hat{O}_{n+1}$ using the PSR Z_n. We train the RNN by minimizing the mean squared loss between the predicted and realized observations, $\hat{O}_{n+1}$ and O_{n+1} respectively. For the campaign we discuss in Section 5, we implement the recurrent module with two gated recurrent unit (GRU;

Cho et al. 2014) layers with 512 and 128 units respectively. Similarly, we implement the predictive module with two fully connected layers with 64 and 29 units respectively.

4.2. Solving the Bellman Equation

Because PSRs satisfy the Markov property, the Bellman equation in Equation 5 can be rewritten with the histories replaced by their corresponding PSRs.

$$Q_{\pi_N}(z, b) = \mathbb{E}\left[\left(r - \rho_{\pi_N}\right) + Q_{\pi_N}(z', b')\right] \quad (7)$$

The Bellman equation holds for every auction. Moreover, at each auction, it only depends on the corresponding PSR z, bid submitted b, the immediate conversion r, and the PSR at the subsequent auction for the same user z'. Therefore, for the purposes of solving the Bellman equation, an auction is completely characterized by the tuple (z, b, r, z'). For the rest of this section, we assume that the observed data is a collection of tuples of the form (z_m, b_m, r_m, z'_m), where $m = 1, \ldots, M$ indexes auctions in some particular order.

In order to satisfy the overlap condition discussed in the Section 3, we augment this list of tuples. Consider a tuple (z, b, r, z') where the focal advertiser lost the auction. If the bid had been zero instead of b, the advertiser would have still observed the same conversion r and the same PSR z' at the subsequent auction. Therefore, for the purposes of learning the Q-function, $(z, 0, r, z')$ is a valid tuple. This observation helps us learn $Q_{\pi_N}(z, 0)$—which informs the estimate of ρ_{π_N} through the Bellman equation—even when the advertiser uses a deterministic algorithm that maps z to a bid b. For every auction in the data that the focal advertiser lost, we augment the data with tuples of the form (z, b', r, z'), where $b' < b$. Adding these additional tuples other than just $(z, 0, r, z')$ helps with efficiency in learning the Q-function.

To solve the Bellman equation, we first represent the Q-function as a neural network, referred to as the Q-network henceforth. Let w denote the vector of all weights of the Q-network.

We denote the Q-function corresponding to an arbitrary w by $\hat{Q}(z, b; w)$. We now present an iterative algorithm that solves the Bellman equation to estimate w_{π_N} and ρ_{π_N}, where $\hat{Q}(z, b; w_{\pi_N}) \approx Q_{\pi_N}(z, b)$.

4.2.1. Temporal Difference Algorithm

For each tuple (z, b, r, z'), we define the temporal difference error or the TD-error (Sutton and Barto 2018) for arbitrary w, ρ as follows

$$\delta_{w,\rho}(z, b, r, z') = (r - \rho) + \hat{Q}(z', 0; w) - \hat{Q}(z, b; w) \tag{8}$$

The TD-error is related to the Bellman equation described in Equation 5. For a given w and ρ, the difference between the left-hand side and the right-hand side of the Bellman equation is known as the Bellman error (Sutton and Barto 2018). It is evident from Equations 7 and 8 that the expectation of the TD-error over all tuples (z, b, r, z') is equal to the Bellman error.

Intuitively, when $w = w_{\pi_N}$ and $\rho = \rho_{\pi_N}$, the expectation of the TD-error must be equal to zero. Therefore, we can implement a stochastic approximation method (Pasupathy and Kim 2011) that iteratively updates w and ρ using the TD-error such that its expectation is equal to zero at convergence. At the start of the iterative process, we arbitrarily initialize w_1 and ρ_1. At each iteration k, we select a tuple (z, b, r, z') uniformly at random from the data. We then compute the TD-error $\delta = \delta_{w_k,\rho_k}(z, b, r, z')$ and update w, ρ using

$$w_{k+1} = w_k - \alpha_w \delta \nabla \hat{Q}(\mathrm{z}, \mathrm{b}; \mathrm{w\_k})$$

$$\rho_{k+1} = \rho_k + \mathbb{I}(b = 0)\alpha_\rho \delta \tag{9}$$

where $\nabla \hat{Q}(z, b; w)$ is the gradient of $\hat{Q}(z, b; w)$ with respect to w; $\mathbb{I}(\cdot)$ is the indicator function; and α_w, α_ρ are the learning rates or step-sizes. We provide the implementation details of the Q-network in Appendix A.

These iterative update equations are equivalent to the Robbins-Monro algorithm, a stochastic approximation method that guarantees convergence to w_{π_N} and ρ_{π_N} (Pasupathy and Kim 2011). More generally, this algorithm belongs to a class of algorithms known as semi-gradient TD methods (Szepesvari 2010). Semi-gradient TD methods approximate a stochastic gradient descent update of a differentiable objective function, known as the projected Bellman error (Liu et al. 2012). Under fairly general conditions, they have been shown to converge robustly, even with highly non-linear classes of functions such as neural networks (Maei et al. 2009). Under convergence, the estimates correspond to the maximum likelihood model of the underlying MDP and are therefore consistent (Sutton and Barto 2018).

5. EMPIRICAL ANALYSIS AND VALIDATION

In this section, we demonstrate how the proposed method can be applied to measure the incremental impact of a campaign. We further validate the estimate using a field experiment.

5.1. Experimental Design

Throughout this section, we use data collected from a campaign run by a collaborating advertiser. The advertiser manages an online store that sells products in the general merchandise category. The campaign was active for four weeks—December 7, 2020 to January 4, 2021—during which users were served impressions through RTB.

The users represent a random sample from the total population of users that satisfy a set of predetermined matching criteria. Each user, identified by a cookie, is randomly assigned to the treatment and control groups with probabilities 0.7 and 0.3 respectively.

The users in the control group were not eligible to receive impressions. The campaign submitted a constant bid of zero for all the auctions corresponding to the control group users. Thus, the ads served to the control group through the auction process are those that would have been served had the campaign not been run. The outcomes of the control group, therefore, provide a valid counterfactual to evaluate the campaign effectiveness. For the treatment group, the campaign calculated bids for each auction using a targeting algorithm. The targeting algorithm took into consideration the browsing history of the user, the number and timing of impressions previously served to the user, and the history of the user's activity on the advertiser's website.

The objective of the campaign was to generate traffic to the online store. Accordingly, a conversion is defined as a session of user activity on the advertiser's website. Following prior research (Jansen et al. 2007) and the industry standard (Ulmer 2010, Google 2021a), we define a browsing session as a continuous period of user activity, where successive events are separated by

no more than 30 minutes. The advertiser observes conversions using a "conversion pixel"—a piece of code embedded in its web pages. The advertiser observes conversions for all users, irrespective of whether they are in the treatment or the control group.

The campaign served four different creatives to the users during its course. These creatives shared a consistent message and varied only slightly in terms of imagery and text. Consequently, we treat these creatives as interchangeable to evaluate advertising effectiveness.

5.2. Data

The data contain information on approximately 286 million RTB auctions for 118,244 users. The bid request for each auction contains the corresponding user's cookie identifier, the timestamp at which the auction had been initiated, and an approximate geographic location of the user. The bid request also contains contextual information that includes the publisher's domain, the URL of the web page on which the impression would be served, the ad exchange's categorization of the web page into one or more verticals (e.g. news, games, shopping), and the size and relative location of the ad space on the web page.

The publishers' web pages are categorized into 26 unique verticals listed in Table 1. Each auction is associated with a 26-dimensional vector, whose elements correspond to the verticals. Each element is a number between 0 and 1, denoting the likelihood that the web page—on which the impression would be served—belongs to the corresponding vertical.

The primary determinant of bids in each auction is provided by the vertical information. At each auction, the targeting algorithm builds a "profile" of verticals unique to the user. The profile measures the user's propensity for browsing pages from each vertical. The targeting algorithm builds two such profiles, one using the entire browsing history, and another for the current browsing session. In addition to the verticals, the targeting algorithm also uses information

on past impressions and conversions—total number and timing of each—to account for pacing and wear-out. Finally, the bid also depends on publishers' historic click-through rate provided by the ad-exchange.

We observe all the determinants of bids in our data. At each auction, the observation O_n contains verticals information of the current page, whether an impression or conversion occurred after the last auction, time elapsed since the last auction, and the historic click-through rate of the current publisher. Thus, each bid is completely determined by the history of observations H_n. As a consequence, the backdoor criterion discussed in Section 3 is satisfied by definition.

Table 2 provides a summary of the data. Of the 118,244 users, 82,612 were assigned to the treatment group, while 35,632 users were assigned to the control group. The campaign served a total of 400,239 impressions—an average of 4.8 impressions per user in the treatment group. The campaign also observed a total of 1,320 conversions—1,101 and 219 respectively from users in the treatment and control groups. It is worth noting that most users who convert do so only once, with only 6 users converting twice.

5.3. RCT Estimate

Here, we compute the incremental impact of the campaign by using data from both the treatment and control groups. Because the users are randomly assigned to the treatment and control groups[1], the causal effect of the campaign is the difference between their average user conversion rates (Imbens and Rubin 2015). The conversion rates for the treatment and control groups are $\frac{1101}{82612} = 1.33\%$ and $\frac{219}{35632} = .61\%$ respectively. Thus, the impact of the campaign is $1.33 - .61 =$

[1] We performed a variety of randomization checks and found no evidence against proper randomization.

.72%. The percentage lift in conversions attributable to the campaign is therefore $\frac{1.33-.61}{.61} =$ 116.8%.The 95% bootstrapped confidence interval[2] for the lift is [89.93%, 145.68%].

In the rest of this section, we use the lift computed here as a benchmark to assess various observational methods, including the proposed MDP framework. The observational methods compute their estimates using data only from the treatment group.

5.4. Proposed Method and Alternatives

We now compute the impact of the campaign using the proposed method and several alternatives. First, we treat the data as cross-sectional and apply the state-of-the-art double/debiased machine learning (DML; Chernozhukov et al. 2018). Consistent with the literature, this approach produces significantly biased estimates. This is because the cross-sectional view of the data does not account for the dynamics of the data generating process—neither the ad serving process nor the effects of impressions. We then account for the dynamic ad serving process by using the PSRs at each auction as control variables. While the resulting bias is smaller, the estimate is still far from that of the RCT. The proposed method, on the other hand, accurately recovers the RCT estimate because it accounts for the dynamics of the ad serving process though the PSRs and the impression effects through the value function.

5.4.1. Cross-sectional View

We conduct the analysis at the user level. Following the extant literature, we define 'treatment' as a binary variable indicating whether a user is exposed to one or more impressions during the campaign period, and the 'outcome' variable as the total number of conversions by that user. All the information obtained through the auctions—verticals and bids submitted—are collected as a vector of 'control' variables. We also include the total number of auctions in the

[2] All the confidence intervals in this section are computed by bootstrapping at the user level.

control variables to account for activity bias. Together, these variables constitute the same information used by the proposed algorithm. To account for the variable number of auctions for each user, we reduce the vector of control variables to a fixed length representation using a standard recursive neural network (Socher 2014). The resulting DAG is shown in Figure 4, where X is the vector of control variables, T is the treatment, and Y is the outcome.

DML (Chernozhukov et al. 2018) provides a non-parametrically efficient estimator for the average treatment effect on the treated (ATT) on such a DAG. It uses machine learning methods to compute the propensity score $\nu(X) = \Pr(T = 1 \mid X)$ and the conditional expectation $\mu(X, T) = \mathbb{E}[Y \mid X, T]$. The ATT is then computed as

$$ATT = \frac{1}{N_e} \sum_{i=1}^{N_u} T_i \left[\frac{T_i\big(Y_i - \mu(X_i, 1)\big)}{\nu(X_i)} - \frac{(1 - T_i)\big(Y_i - \mu(X_i, 0)\big)}{1 - \nu(X_i)} + \mu(X_i, 1) - \mu(X_i, 0) \right]$$

where N_u is the total number of users and N_e is the number of users treated. The resulting estimate is root-n consistent and non-parametrically efficient (Chernozhukov et al. 2018). Moreover, its mean squared error outperforms that of traditional methods like regression adjustment, inverse probability weighting, and propensity score matching. Because $\nu(X)$ and $\mu(X, T)$ are non-parametrically estimated, DML can be applied to settings with a large number of control variables (compared to sample size) by using methods such as regularization.

Note that the ATT provided by DML is not directly comparable to the intent to treat (ITT) estimate provided by the RCT. The ITT, however, can be computed by simply multiplying the ATT with the fraction of users treated in the data (Gordon et al. 2019). The percentage lift is then computed as

$$lift = \frac{ITT}{\eta_u - ITT} \times 100 \tag{10}$$

where η_u is the average number of conversions observed per unit.

If the DAG in Figure 4 reflects the true data generating process, we expect DML to recover the RCT estimate. When we apply DML to our data, however, the resulting lift is equal to 184.92% with a confidence interval of [175.71%, 194.29%], overestimating the RCT lift by 68.12%. This large bias is a result of an inaccurate model of the data generating process. When collapsed to a cross-section, information on the mapping from history of observables to the probability of exposure as well as the timing and interactive effects of impressions is lost. As a result, approaches that adopt the cross-sectional view cannot account for the dynamic nature of the ad serving process nor the dynamic effects of impressions.

5.4.2. Dynamic Ad Serving

Here, we account for one of two sources of dynamics in the data generating process—the ad serving mechanism. To do so, we conduct the analysis at the auction level. At each auction, we compute the PSR as described in Section 4.1. Because the PSR is a representative summary of the history of observations, it accounts for selection due to targeting by the focal advertiser. We use the PSR as a set of 'control' variables, whether an impression is served as the 'treatment', and an indicator for conversion before the subsequent auction as the 'outcome' variable.

We construct the PSRs using the PSR-network that takes as input observations O_n. At each auction n, the observation O_n is a 29-dimensional vector. The first element of O_n is the time—measured in seconds—since the preceding auction. This allows the model to compute any lagged effects that depend on time. The next 26 elements represent the vertical information of the publisher's webpage, each specifying the likelihood—between 0 and 1—that the webpage belongs to its corresponding vertical. The penultimate element of O_n is an indicator for whether the focal advertiser's impression is served through auction n. The last element of O_n is an indicator for whether a conversion occurred between auctions $n-1$ and n.

Similar to Section 5.4.1, we implement DML to compute the percentage lift in conversions attributable to the campaign. The resulting lift is 158.65% with a confidence interval of [148.42%, 169.39%]. Although this represents an improvement over the cross-sectional analysis, the resulting error of 41.85% is still large. This is a result of analyzing each auction independently. An impression's effect only on the immediate conversion is considered, ignoring its temporal effects and interactive effects with other impressions served to the same user. Note that these dynamic effects of impressions are accounted for by the Q- function in the proposed method.

5.4.3. Proposed Method

Here, we estimate the impact of the campaign using the proposed MDP framework. The advertiser interacts with any given user through a sequence of RTB auctions indexed by n. The estimation procedure discussed in Section 4 depends only on the sequence of advertiser's observations $\{O_1, O_2, \dots\}$, bids $\{B_1, B_2, \dots\}$, and conversions $\{R_1, R_2, \dots\}$.

To estimate the impact of the campaign, we compute the average conversion rates per auction ρ_{π_A} and ρ_{π_N} corresponding, respectively, to the advertiser's policy π_A and the null policy π_N. ρ_{π_A} can be computed from Table 2 as the ratio of total number of conversions observed and the total number of auctions, $\rho_{\pi_A} = \frac{1101}{200498908} = 5.49 \times 10^{-6}$. Using the proposed method, we estimate ρ_{π_N} to be 2.47×10^{-6}. The estimated lift is thus $\frac{5.49-2.47}{2.47} = 122.19\%$. The 95% bootstrapped confidence interval for the estimated lift is [108.3%, 136.19%].

Thus, the proposed method's estimate of the impact of the campaign is within 6% of the RCT estimate and lies within the confidence interval of the RCT estimate. It is also worth noting that the difference between the two estimates is about .31 times the standard error of the RCT estimate. This shows that the proposed method can accurately measure the effects of display advertising. Moreover, this shows that observational approaches can be used for ad measurement

if the different components of the ad serving process—advertiser's decision making, the user's response to impressions, and the advertiser's total payoff—are explicitly modeled.

6. CONCLUSION

In this paper, we propose an observational framework to measure the incremental impact of display advertising campaigns. The framework accounts for the different sources of endogeneity present in the RTB ecosystem by explicitly modeling the users' browsing behavior, the advertiser's decision making, and the users' response to impressions. We leverage techniques from the reinforcement learning literature to develop a novel estimation method that accurately recovers counterfactual outcomes. Empirically, we validate the estimate provided by the proposed method through an RCT. We show that the proposed method's estimate of the impact of a campaign is within 6% of the RCT estimate.

Our framework contributes to the growing literature on measuring display advertising effectiveness. Gordon et al. (2019) and Lewis et al. (2011) show that commonly used observational methods produce significantly biased estimates. As a result, the recent approaches have focused on integrating experimentation directly into the advertiser's targeting algorithms (Gordon et al. 2021). However, we show here that observational data can indeed be used to assess the impact of display advertising. Thus, we view our framework as a useful addition to the literature, and we believe that it could play a valuable role in the display advertising landscape. We also contribute to the rapidly growing literature at the intersection of causal inference and machine learning by showing that techniques from reinforcement learning can be applied for inference in settings that involve sequential decision making such as display advertising.

Leveraging our approach offers several advantages. Because the approach uses readily available observational data, advertisers can implement it without incurring the opportunity cost associated with experimentation. Furthermore, our approach allows researchers and advertisers to explore moderating factors of the impact of impressions, providing insights into potential changes

in ad responsiveness resulting from inter-temporal shifts in a user's state. This information enables advertisers to examine the impacts of different targeting strategies and to optimize campaign performance.

Future research can extend this work in several ways. First, although reinforcement learning methods perform well in practice, little is known about their asymptotic efficiency. Future research should explore how the variance of the proposed method's estimate depends on the sample size, campaign attributes, and various neural network architectures. Similar to Gordon et al. (2019), future research could also investigate the empirical performance of the proposed method and compare it to RCTs over a large number of campaigns. Second, the proposed method relies on the assumption that the advertiser's bid and the user state at any auction are unconfounded conditional on the history of observations. While this assumption holds in the display advertising context where bid decisions are made algorithmically, in settings such as those involving salesforces, the researcher does not always observe all the inputs to the agent's decision making. This highlights the need for future research to extend the proposed method to accommodate techniques such as instrumental variables. Finally, future research could extend the MDP framework to recover an optimal targeting policy while satisfying advertiser constraints such as budgets. The extant literature on reinforcement learning mainly focuses on settings where a single criterion is to be maximized. However, in the RTB ecosystem, budget constrained advertisers seek to maximize conversions while keeping the total cost below the budget.

REFERENCES

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866.

Awad MA, Khalil I (2012) Prediction of user's web-browsing behavior: Application of Markov model. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(4):1131–1142

Balachandran V, Deshmukh S (1976) A stochastic model of persuasive communication. Management Science 22(8):829–840.

Bowling M, Burch N, Johanson M, Tammelin O (2017) Heads-up limit hold'em poker is solved. Communications of the ACM 60(11):81–88, ISSN 00010782.

Braun M, Moe WW (2013) Online display advertising: Modeling the effects of multiple creatives and individual impression histories. Marketing Science 32(5):753–767.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters.

Cho K, Van Merri¨enboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Choi H, Mela CF, Balseiro SR, Leary A (2020) Online display advertising markets: A literature review and future directions. Information Systems Research 31(2):556–575.

Deng Y, Bao F, Kong Y, Ren Z, Dai Q (2016) Deep direct reinforcement learning for financial signal representation and trading. IEEE transactions on neural networks and learning systems 28(3):653–664.

Downey C, Hefny A, Gordon G (2017a) Practical learning of predictive state representations. arXiv preprint arXiv:1702.04121.

Downey C, Hefny A, Li B, Boots B, Gordon G (2017b) Predictive state recurrent neural networks. arXiv preprint arXiv:1705.09353.

Glavic M, Fonteneau R, Ernst D (2017) Reinforcement learning for electric power system decision and control: Past considerations and perspectives. IFAC-PapersOnLine 50(1):6918–6927.

Google (2021a) How a web session is defined in universal analytics. Website, URL https://support.google. com/analytics/answer/2731565, accessed 2021-02-08.

Google (2021b) Real time bidding: Latency restrictions and peering. Website, URL https://developers. google.com/authorized-buyers/rtb/peer-guide, accessed 2021-02-16.

Gordon BR, Jerath K, Katona Z, Narayanan S, Shin J, Wilbur KC (2021) Inefficiencies in digital advertising markets. Journal of Marketing 85(1):7–25.

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. Marketing Science 38(2):193–225.

Hauser JR, Wisniewski KJ (1982) Dynamic analysis of consumer response to marketing strategies. Management Science 28(5):455–486.

Hausknecht M, Stone P (2015) Deep recurrent q-learning for partially observable mdps. arXiv preprint arXiv:1507.06527.

Heckman JJ, Pinto R (2003) Causality and econometrics. University of Chicago and the American Bar Foundation, 2003 Draft.

Hoban PR, Bucklin RE (2015) Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. Journal of Marketing Research 52(3):375–393.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. The Review of Economic Studies 60(3):497–529.

IAB (2021) 2020/2021 iab internet advertising revenue report. Website, URL https://www.iab.com/ insights/internet-advertising-revenue-report/, accessed 2022-02-01.

Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences (Cambridge University Press).

Jansen BJ, Spink A, Blakely C, Koshman S (2007) Defining a session on web search engines. Journal of the American Society for Information Science and Technology 58(6):862–871.

Johnson GA, Lewis RA, Nubbemeyer EI (2017) Ghost ads: Improving the economics of measuring online ad effectiveness. Journal of Marketing Research 54(6):867–884.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research 17(1):1334–1373.

Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D (2018) Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research 37(4-5):421–436.

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics 130(4):1941–1973.

Lewis RA, Rao JM, Reiley DH (2011) Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. Proceedings of the 20th international conference on World wide web, 157–166.

Li Y, Liang X, Hu Z, Xing EP (2018) Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in Neural Information Processing Systems, 1530–1540.

Littman ML, Sutton RS, Singh SP (2001) Predictive representations of state. NIPS, volume 14, 30.

Liu B, Mahadevan S, Liu J (2012) Regularized off-policy td-learning. Advances in Neural Information Processing Systems 25:836–844.

Maei HR, Szepesvari C, Bhatnagar S, Precup D, Silver D, Sutton RS (2009) Convergent temporal-difference learning with arbitrary smooth function approximation. NIPS, 1204–1212.

Moody J, Saffell M (2001) Learning to trade via direct reinforcement. IEEE transactions on neural Networks 12(4):875–889.

Pasupathy R, Kim S (2011) The stochastic root-finding problem: overview, solutions, and open questions.

ACM Transactions on Modeling and Computer Simulation (TOMACS) 21(3):1–23.

Pearl J (2000) Models, reasoning and inference. Cambridge, UK: Cambridge University Press 19:2.

Peng YS, Tang KF, Lin HT, Chang E (2018) Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. Advances in Neural Information Processing Systems, 7322–7331.

Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming (John Wiley & Sons).

Reiss PC, Wolak FA (2007) Structural econometric modeling: Rationales and examples from industrial organization. Handbook of econometrics 6:4277–4415.

Robins JM (2004) Optimal structural nested models for optimal sequential decisions. Proceedings of the second seattle Symposium in Biostatistics, 189–326 (Springer).

Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 100(469):322–331.

Rust J (1994) Structural estimation of markov decision processes. Handbook of econometrics 4:3081–3143.

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

Schwartz A (1993) A reinforcement learning method for maximizing undiscounted rewards. Proceedings of the tenth international conference on machine learning, volume 298, 298–305.

Scott SL, Hann IH (2006) A nested hidden markov model for internet browsing behavior. Marshall School of Business 1–26.

Shen J, Orten B, Geyik SC, Liu D, Shariat S, Bian F, Dasdan A (2015) From 0.5 million to 2.5 million: Efficiently scaling up real-time bidding. 2015 IEEE International Conference on Data Mining, 973–978 (IEEE).

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. (2017) Mastering the game of go without human knowledge. Nature 550(7676):354.

Singh S, James M, Rudary M (2012) Predictive state representations: A new theory for modeling dynamical systems. arXiv preprint arXiv:1207.4167.

Socher R (2014) Recursive deep learning for natural language processing and computer vision (Stanford University).

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction (MIT press).

Szepesv´ari C (2010) Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning 4(1):1–103.

Thon MR, Jaeger H (2015) Links between multiplicity automata, observable operator models and predictive state representations: a unified learning framework. J. Mach. Learn. Res. 16:103–147.

Tsitsiklis JN, Van Roy B (2001) Regression methods for pricing complex american-style options. IEEE Transactions on Neural Networks 12(4):694–703.

Tunuguntla S, Hoban PR (2021). A near-optimal bidding strategy for real-time display advertising auctions. Journal of Marketing Research, 58(1), 1-21.

Ulmer H (2010) Browsing sessions. Website, URL https://blog.mozilla.org/metrics/2010/12/22/browsing-sessions/, accessed 2021-02-08.

Upadhyay U, De A, Rodriguez MG (2018) Deep reinforcement learning of marked temporal point processes. Advances in Neural Information Processing Systems, 3168–3178.

Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. Proceedings of the AAAI Conference on Artificial Intelligence, volume 30.

Visser I (2011) Seven things to remember about hidden markov models: A tutorial on markovian models for time series. Journal of Mathematical Psychology 55(6):403–415.

Visser I, Raijmakers ME, Molenaar P (2002) Fitting hidden markov models to psychological data. Scientific Programming 10(3):185–199.

Wickens TD (1982) Models for behavior: Stochastic processes in psychology (WH Freeman & Co Ltd). Zantedeschi D, Feit EM, Bradlow ET (2017) Measuring multichannel advertising response. Management Science 63(8):2706–2728.

Zhu Y, Simester D, Parker JA, Schoar A (2020) Dynamic marketing policies: Constructing markov states for reinforcement learning. Available at SSRN 3633870.

Appendix A. ALGORITHM IMPLEMENTATION

For any given counterfactual policy π, we estimate the corresponding Q-function Q_π and the average conversion rate ρ_π in two steps. First, we use the sequence of advertiser observations $\{O_1, O_2, \dots\}$ and bids $\{B_1, B_2, \dots\}$ to train a recurrent neural network (RNN) that generates a sequence of predictive state representations (PSRs) $\{Z_1, Z_2, \dots\}$. Next, we apply a temporal difference algorithm to the sequence of PSRs $\{Z_1, Z_2, \dots\}$, bids $\{B_1, B_2, \dots\}$, and rewards $\{R_1, R_2, \dots\}$ to train the Q-network that simultaneously estimates the Q-function Q_π and the average conversion rate ρ_π.

A.1. PSR Network

The model architecture is shown in Figure 3. The model has two key modules: a recurrent module that computes the PSRs, and a predictive module that predicts subsequent observations using the PSRs.

The recurrent module consists of 2 layers with 512 and 128 gated recurrent units (GRUs; Cho et al. 2014) respectively. A GRU layer, which is one variant of an RNN, processes a sequence of inputs $\{x_1, \dots, x_N\}$ to compute a sequence of "internal states" $\{h_1, \dots, h_N\}$, where x_n, h_n are vectors of fixed dimensions. For each n, the GRU layer uses h_{n-1} and x_n to compute h_n as follows

$$h_n = p_n \odot h_n + (1 - p_n) \odot sigmoid(W_h x_n + U_h[q_n \odot h_{n-1}])$$

Where W_h, U_h are matrices representing the parameters of the layer, and $\odot$ represents element-wise multiplication of vectors. p_n, q_n—referred to as update and reset gates respectively—are computed using

$$p_n = sigmoid(W_p x_n + U_p h_{n-1})$$

$$q_n = sigmoid(W_q x_n + U_q h_{n-1})$$

The initial internal state h_0 is set equal to a vector of zeros. For a more general discussion of GRUs, see Cho et al. (2014).

The first GRU layer of the recurrent module takes as input the sequence of tuples (O_n, B_n) to compute its sequence of internal states. The output of the first layer is fed as input to the second layer. The sequence of internal states of the final layer in the recurrent module is the sequence of PSRs $\{Z_1, Z_2, \ldots\}$.

The predictive module, at each auction n, uses the PSR Z_n to predict the subsequent observation $\hat{O}_{n+1}$. It consists of 2 fully connected layers with 64 and 29 units respectively.

We train the recurrent and predictive modules simultaneously by minimizing the mean squared error between the predicted and realized observations. We split our dataset into training and validation sets containing 80% and 20% of the users respectively. Using the training set, we train our model using the Adam algorithm (Kingma and Ba 2014), a variant of stochastic gradient descent that is well suited for problems that are large in terms of data and parameters. The hyper-parameters of the model and training process such as the number of hidden units, number of layers, and learning rate are selected such that they minimize the mean squared error over the validation set.

A.2. Q-Network

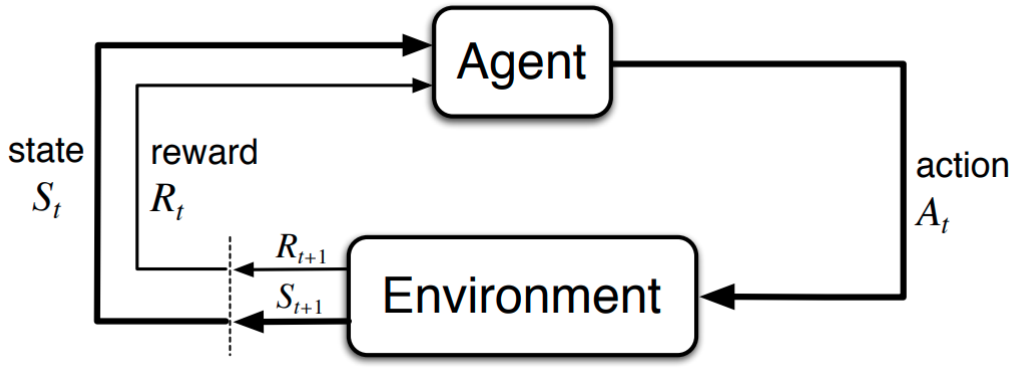
The Q-network estimates the Q-function $Q_\pi(z, b)$ where z is a PSR and b is a bid amount. The Q-network consists of 3 fully connected hidden layers with 1024, 256, and 1 unit respectively.

Each auction is denoted by a tuple (z, b, r, z'), where z is the PSR at the auction, b is the bid, r is an indicator for a conversion between the auction and the subsequent auction, z' is the PSR at the subsequent auction.

While the iterative update algorithm described by Equation 9 can be implemented directly, to ensure numerical stability, we use two identical Q-networks: a primary and a target network (Van Hasselt et al. 2016). The TD-error in Equation 9 is computed using the target network, while the gradient is computed with respect to the primary network. The parameters of the primary network are updated at each iteration, while the parameters of the target network are set equal to those of the primary network every 1000 iterations. This approach has been shown to increase numerical stability (Van Hasselt et al. 2016) and is widely used in reinforcement learning algorithms (Hausknecht and Stone 2015).

Similar to the PSR network's training procedure, we split our dataset into training and validation sets containing 80% and 20% of the auctions respectively. The temporal difference algorithm is applied using the training set. The hyper-parameters of the model and training process are selected such that the Bellman error is minimized over the validation set.

Figure 1. Markov Decision Processes



Note. The agent-environment interaction in a Markov decision process. Reprinted from Sutton and Barto (2018)

Figure 2. Data Generating Process

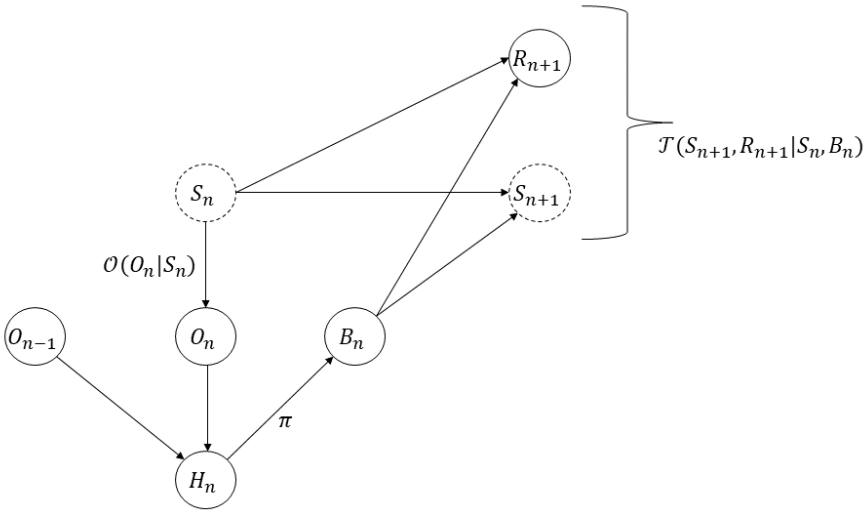


Figure 3. PSR Network – Model Architecture

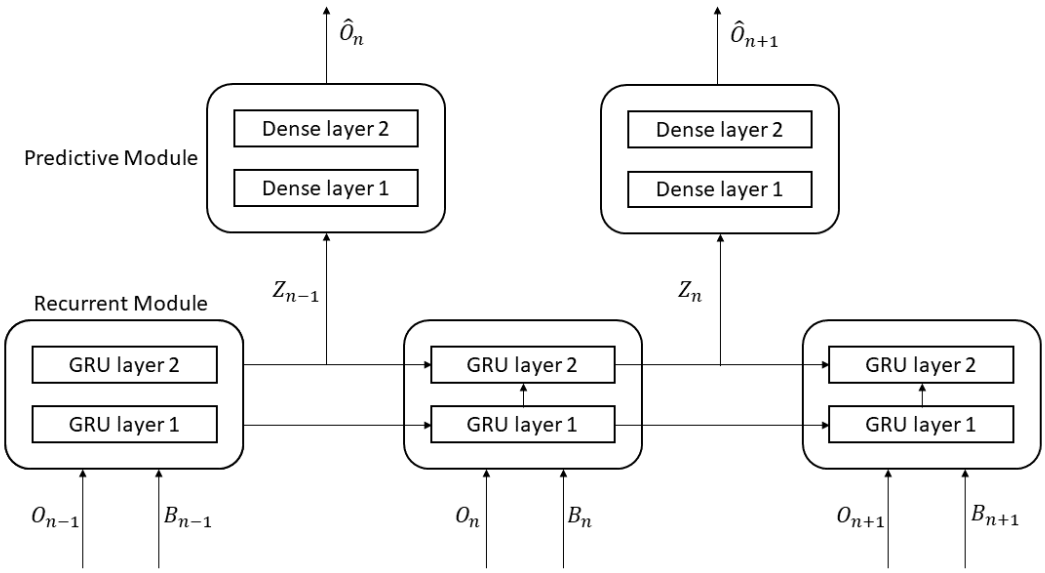


Figure 4. DAG for Cross-sectional View of the Data

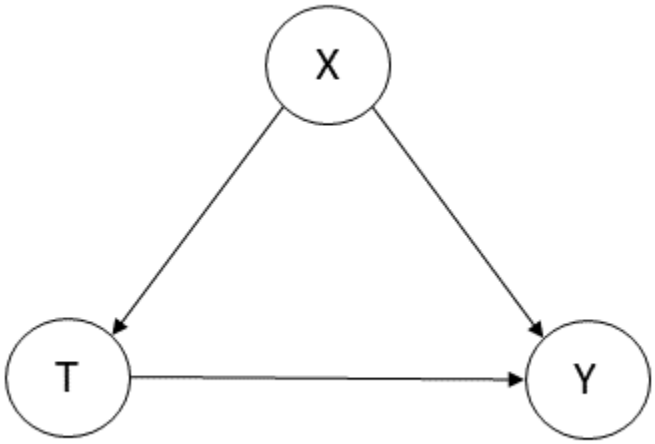


Table 1. List of Verticals

Arts & Entertainment	Health	Pets & Animals
Autos & Vehicles	Hobbies & Leisure	Real Estate
Beauty & Fitness	Home & Garden	Reference
Books & Literature	Internet & Telecom	Science
Business & Industrial	Jobs & Education	Shopping
Computers & Electronics	Law & Government	Sports
Finance	News	Travel & Transportation
Food & Drink	Online Communities	World Localities
Games	People & Society	

Table 2. Summary of Campaign Data

	Treatment	Control	Total
Users	82,612	35,632	118,244
Auctions	200,498,908	86,327,804	286,826,712
Impressions	400,239	0	400,239
Conversions	1101	219	1320