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 because of small effect sizes and the selection biases induced by targeting. The proposed framework, derived from Markov Decision Processes, 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 (1) a display advertising campaign and (2) the individual impressions served. We validate the proposed method using a randomized controlled trial.

Key words: 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 in targeting as the key driver of return on investment (Forbes 2015).

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 the advances in targeting, such causal measurement remains a challenging issue. While randomized controlled trials (RCTs) provide an unbiased estimate of the average effect on the exposed users, firms are often unable or unwilling to conduct RCTs due to considerations of cost and lack of infrastructure (Gordon et al. 2021). Observational methods, on the other hand, face challenges because: (1) the incremental impact of the impressions is small relative to the volatility of the outcomes of interest, i.e., only a small amount of variation in outcomes is explained by ad exposures, (2) individual level outcomes are highly heterogeneous, and (3) the targeted nature of advertising—the non-random assignment of exposures—induces endogeneity. Lewis et al. (2011) and Gordon et al. (2019) show that commonly used observational methods produce significantly biased estimates of the impact of digital advertising.

In this paper, we propose a framework to measure the incremental impact of a campaign, as well as the impact of individual impressions. The framework uses readily available observational data, makes few assumptions on the data generating process, and accounts
for selection biases and heterogeneity in individual outcomes to recover the incremental lift in outcomes.

We model the data generating process of an ad campaign as a Markov Decision Process (MDP). In the model, each user is described by a *state* that evolves over time as a Markov process. Visser et al. (2002) and Awad and Khalil (2012) find that psychological processes, especially online browsing behaviors, are well described by Markov processes. Markov states correspond to the cognitive and emotional states, while state transition probabilities describe the evolution of these states. Any form of persuasive communication, including advertisements, alter the user state (Balachandran and Deshmukh 1976, Hauser and Wisniewski 1982). Accordingly, we model the advertiser as an agent driving conversions by inducing alterations to the state that influence the user to take action.

The MDP framework provides a way to explicitly model (1) the advertiser’s decision making—as a mapping from the information available on the user to an intent to serve an impression, (2) the users’ response to impressions—via the state transition probabilities, and (3) the advertiser’s payoff—as a mapping from the user state to a probability of conversion. Furthermore, the user’s response to an impression—the state transition probabilities as a function of exposure—is invariant to the advertiser’s targeting policy. In other words, the state transition probabilities are *structural* parameters (Reiss and Wolak 2007). Thus, any counterfactual outcomes that depend only on these structural parameters can be accurately estimated using observational data. As we show in Section 3.2, the counterfactual outcomes had the users not seen any ads depend only on the state transition probabilities.

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 algorithm. 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.

Furthermore, we show that the commonly used observational methods provide significantly biased estimates even when the observables satisfy the unconfoundedness assumption, i.e., when we observe all the inputs to the advertiser’s decision making. This is likely because the commonly used methods like regression adjustment are not sufficiently expressive to model the dynamic nature of RTB data and they further make strong assumptions on the data generating process (e.g., a particular form of outcome or exposure model). Thus, we show that observational methods that use flexible statistical descriptions of the data and explicitly model the strategic behavior of the advertiser can be used to accurately measure the effects of display advertising.

We contribute to the display advertising literature by providing a novel method that accurately estimates the causal impact of advertising. The method uses readily available observational data, avoiding the large costs associated with RCTs. Furthermore, we measure the impact of individual impressions, providing insights into potential changes in ad responsiveness resulting from inter-temporal shifts in a user’s state. Leveraging these potential differences in ad responsiveness is critical to maximizing campaign success. 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 settings that involve sequential decision making such as display advertising.
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 Section 3, we present the MDP framework and an estimation method to recover counterfactual outcomes. Section 4 describes the field experiment, validates the estimate provided by the proposed method, and computes the impact of individual impressions. Section 5 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.

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 or an impression 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 from 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.

Experimental approaches also provide little guidance on how to improve campaign impact. Specifically, they are unable to measure the impact of individual impressions which allows for investigation of moderating factors such as the timing and volume of impressions, and the diminishing returns of impressions.

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, an observational method that sufficiently models the data generating process can provide an accurate estimate of the return on investment.

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

In the absence of randomized assignment of exposures, observational approaches model the assignment mechanism and the outcome distribution. They vary on the degree of structure they impose on the data generating process. Examples of observational approaches include propensity score matching, instrumental variable methods, inverse probability weighting, and regression adjustment.

Because the exposures are endogenous, failure to control for the assignment mechanism can severely bias the estimates. In display advertising, several sources of endogeneity have been identified that induce systematic differences between the exposed and non-exposed users.

One 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. Thus, approaches that do not account for the targeting criteria tend to overestimate the impact of advertising.
Another source of endogeneity is the competitive effects of ad serving. Because the impressions are sold in auctions, 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). That is, users with higher probabilities of conversion are more likely to be exposed. Similar to the targeting induced endogeneity, failure to account for competitive effects can lead to overestimation of advertising effectiveness. Gordon et al. (2019) argue that a case of underestimation can also occur when other advertisers selling complementary products target the same users as the focal advertiser. If these firms win impressions at the expense of the focal advertiser, the resulting set of unexposed users may be more likely to convert for the competing firms.

Activity bias, first identified by Lewis et al. (2011), is also a mechanism that induces endogeneity. It results from the correlation between different activities users undertake online. Users who browse more intensely are more likely to receive an impression merely because they are likely to be online during the campaign. Lewis et al. (2011) show that they are also more likely to convert merely due to the positive correlation between browsing behaviors across websites. Thus, observational approaches must control for browsing intensity to accurately measure advertising effectiveness.

Finally, observational approaches suffer from inductive bias. Inductive bias results from the additional assumptions required to justify the inductive inferences of a model as deductive inferences (Mitchell 1997). These assumptions define the relationships between the variables of interest and the structure of the data generating process. If the models of assignment of exposures and the outcome distribution are not sufficiently expressive, they can severely bias the estimates (Mitchell 1997). Commonly used observational methods, for
example, use Rubin’s causal model (Rubin 2005) which assumes a static model of assignment. They ignore the dynamic sequential nature of interactions between the advertiser and the users. They control for whether a user was exposed, but not the timing or volume of the exposures.

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 accounts for the different sources of bias. The MDP framework explicitly models the user activity and the advertiser’s decision making, thus controlling for activity bias and targeting induced endogeneity respectively. Competition induced endogeneity occurs because a user’s exposure depends not only on the focal advertiser’s bid but also on other advertisers’ bids. The MDP approach eliminates such competition induced endogeneity by quantifying the impact of bids on conversions rather than the impact of impressions. Finally, our approach reduces inductive bias by explicitly modeling the data generating process, remaining agnostic to the probability distributions characterizing the MDP, and using universal approximators (Cybenko 1989) to model the advertiser’s decision making and outcomes.
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 $O(O_t|S_t)$ 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 $H_t = \{O_1, O_2, \ldots, O_t\}$. 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 $\mathcal{T}(S_{t+1}, R_{t+1}|S_t, A_t)$.

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 $\pi$ 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 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 a 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. Method

3.1. MDP Framework

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 graph in Figure 2. Each node in the graph represents a variable and the edges capture the conditional dependencies among 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 $O(O_n|S_n)$.

The advertiser’s bid $B_n$ typically depends on the entire history of observations denoted by $H_n = \{O_1, O_2, \ldots, O_n\}$. The advertiser’s decision making is characterized by a policy $\pi : H \rightarrow \mathbb{R}$ that maps a history of observations to a bid.

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 $T(S_{n+1}, R_{n+1}|S_n, B_n)$.

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. Because each user is associated with a different number of auctions, we use the average conversions per auction as a measure of the advertiser’s payoff. Thus, the quality of the advertiser’s policy is measured through the average number of conversions obtained per auction. We compute the causal impact of the advertiser’s policy by estimating the average conversions per auction had the advertiser not implemented a campaign.

We assume that the distributions characterizing the MDP—transition and conversion dynamics $T$ and distribution of observations $O$—are unknown. To compute counterfactual outcomes, we present an estimation method that is agnostic to these distributions and depends only on the sequence of advertiser’s observations $\{O_1, O_2, \ldots\}$, bids $\{B_1, B_2, \ldots\}$, and conversion indicators $\{R_1, R_2, \ldots\}$.

3.2. Definitions and Assumptions

Let $\pi_A$ be the advertiser’s policy. To estimate the causal impact of $\pi_A$, we must compute the counterfactual conversions in the absence of the campaign. We capture this counterfactual
scenario by defining a null policy \( \pi_N \). 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 \( \pi \) is associated with a long-term average conversion rate \( \rho_\pi \) defined as

\[
\rho_\pi = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} \left[ R_n \mid B_{0:n-1} \sim \pi \right] = \mathbb{E} \left[ \sum_{k=1}^{\infty} (R_{n+k} - \rho_\pi) \right] \mid H_n = h, B_n = b
\]  

where the expectations are conditioned on bids submitted according to policy \( \pi \). The average rate of conversions \( \rho_\pi \) depends on the steady state distribution of the Markov process while following \( \pi \). Under the assumption of ergodicity, the steady state distribution always exists and is independent of the initial state \( S_1 \) (Mahadevan 1996). Ergodicity is sufficient to guarantee the existence of the limits in the equation (Mahadevan 1996).

The incremental impact of a campaign can then be computed by comparing the average conversion rates resulting from the advertiser’s policy \( \pi_A \) and the null policy \( \pi_N \). For ease of exposition, we express the impact of a campaign as the percentage lift in the conversion rate.

\[
\text{lift}(\pi_A) = \frac{\rho_{\pi_A} - \rho_{\pi_N}}{\rho_{\pi_N}} \times 100
\]

(2)

In addition to measuring the impact of the campaign, we also measure the impact of individual impressions. Recall, the advertiser computes a bid in any given auction based on the history of observations from the user. We define \( Q_\pi(h, b) \) as a measure of the effect of an individual bid \( b \) when the history is \( h \) while following policy \( \pi \).

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

(3)
The function $Q_\pi(h, b)$ is referred to as the state-action value function or the Q-function in the reinforcement learning literature (Sutton and Barto 2018). It measures the incremental impact of an individual bid on the total conversions over the steady state average rate of conversions while following policy $\pi$. Thus, the effect of a counterfactual bid $b'$ can be computed using $Q_\pi(h, b')$.

Consider an auction $n$ following which an impression is served. Let $B_n$ be the bid submitted. The impact of the impression can be estimated by computing the difference $Q_\pi(H_n, B_n) - Q_\pi(H_n, 0)$. The term $Q_\pi(H_n, 0)$ corresponds to the counterfactual bid of zero, in which case an impression would not have been served.

MDPs provide sufficient “structure” for valid inference. For any given policy $\pi$, the average conversion rate $\rho_\pi$ and the Q-function $Q_\pi$ depend only on the transition dynamics of the MDP. This is because a user’s response to an impression is independent of the targeting and bidding algorithms used to make the ad serving decision. Therefore, these transition dynamics are invariant to the policy implemented (Rust 1994, Hotz and Miller 1993). In other words, they are structural parameters with respect to the advertiser’s policy. Thus, any method that estimates the transition dynamics, implicitly or explicitly, can accurately construct counterfactuals. In Section 3.3, we present one such method that simultaneously estimates $\rho_\pi$ and $Q_\pi$, for any given policy $\pi$.

Our approach relies on several standard assumptions for valid inference. The first is similar to the stable unit treatment value assumption in Rubin’s causal model (Rubin 2005). We assume that the exposures and outcomes of one user do not affect any other user’s exposures or outcomes.

Second, we assume that the MDP is ergodic. An MDP is ergodic if the associated Markov chain is ergodic for every deterministic policy (Mahadevan 1996). In particular, we assume
that the transition dynamics of the MDP are stationary. To satisfy this assumption, it is sufficient that the market conditions—such as the focal advertiser’s strategies in other marketing channels and competitors’ marketing strategies—are fixed during the course of the campaign. Gordon et al. (2019) point out that this assumption is reasonable because campaigns are not pre-announced and occur over relatively short periods.

Finally, we assume that at each auction, the bid submitted is unconfounded with the user state conditional on the history of observations. That is, we observe all the inputs to the advertiser’s bid decision. This is akin to satisfying the unconfoundedness assumption in Rubin’s causal model—the bid is analogous to treatment and the user state is analogous to the potential outcomes. Because bid decisions are made algorithmically in display advertising, it is common for advertisers to observe all the inputs to the decision, satisfying this assumption.

3.3. Estimation

For any given policy $\pi$—such as the null policy $\pi_N$—we estimate the corresponding Q-function $Q_\pi$ and the average conversion rate $\rho_\pi$ simultaneously. The estimation method presented here uses only the data obtained from implementing the advertiser’s policy $\pi_A$. That is, it depends only on the sequence of advertiser’s observations $\{O_1, O_2, \ldots\}$, bids $\{B_1, B_2, \ldots\}$, and conversion indicators $\{R_1, R_2, \ldots\}$.

3.3.1. Bellman Equation. In our estimation, we use the property that the Q-function satisfies a Bellman equation (Schwartz 1993). The recursive relationship defined by the Bellman equation can be derived from Equation 3. The right hand side of Equation 3 can be written as

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

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

(4)

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.

3.3.2. Parameterization. Because the space of histories and bids is infinite and continuous, we represent the Q-function as a parameterized functional form with parameters $w \in \mathbb{R}^p$. This parameterized functional form represents a class of functions, where each member of the class corresponds to a value $w$ can take. This class of functions, for example, can be linear functions with $w$ as the vector of linear coefficients, or weighted combination of kernels with $w$ parameterizing the kernel.

In this work, we represent the Q-function as a deep neural network—referred to as the Q-network henceforth—where $w$ denotes the connection weights in all the layers. Neural networks are widely used to construct Q-functions in the reinforcement learning literature (Arulkumaran et al. 2017). They are responsible for the state-of-the-art performance in a wide range of applications (Li 2017). This is because of their universal approximation property: neural networks can approximate any continuous function to an arbitrary degree of approximation (Cybenko 1989).

We denote a function corresponding to an arbitrary parameter vector $w$ by $\hat{Q}(h, b; w)$. For a given policy $\pi$, the goal of the estimation method is to find the vector $w_\pi$ such that $\hat{Q}(h, b; w_\pi) \approx Q_\pi(h, b)$. 
3.3.3. Temporal Difference Algorithm. Here, we present an overview of an iterative algorithm to jointly estimate $w_\pi$ and $\rho_\pi$. We provide the implementation details of the algorithm including neural network architectures, model selection criteria, and techniques to ensure numerical stability in Web Appendix A.

First we represent each auction $n$ by a tuple $(h, b, r, o)$, where $h$ is the history of observations until the auction, $b$ is the bid submitted, $r$ is the indicator for user conversion between auctions $n$ and $n+1$, and $o$ is the observation at the subsequent auction $n+1$.

Next, for each tuple $(h, b, r, o)$, we define the temporal difference error or the TD-error (Sutton and Barto 2018) for arbitrary $w, \rho$ as follows

$$\delta_{w,\rho}(h, b, r, o) = (r - \rho) + \hat{Q}(h', \pi(h'); w) - \hat{Q}(h, b; w)$$

where $h'$ is the history $h$ appended with the observation $o$. The TD-error is related to the Bellman equation described in Equation 4. 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 4 and 5 that the expectation of the TD-error is the Bellman error. We use the TD-error for each tuple $(h, b, r, o)$ to iteratively solve the Bellman equation.

At the start of the iterative process, we arbitrarily initialize $w_1$ and $\rho_1$. At each iteration $k = 1, 2, \ldots$, we select a tuple $(h, b, r, o)$ uniformly at random from the data. We then compute the TD-error $\delta_{w,\rho}(h, b, r, o)$ and update $w, \rho$ using

$$w_{k+1} = w_k - \alpha_w \delta_{w,\rho} \nabla \hat{Q}(h, b; w_k)$$
$$\rho_{k+1} = \rho_k + \alpha_\rho \delta_{w,\rho}$$

where $\nabla \hat{Q}(h, b; w)$ is the gradient of $\hat{Q}(h, b; w)$ with respect to $w$. The learning rates $\alpha_w, \alpha_\rho$ are hyper-parameters that are selected using a hold-out validation set, as described in Web Appendix A.
Theoretically, these iterative update equations belong to a class of algorithms known as semi-gradient TD methods (Szepesvári 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 (Sutton and Barto 2018). Thus, they are consistent and asymptotically normal.

3.3.4. Predictive State Representation. While the algorithm presented in the previous subsection can be implemented in theory, there is a potential computational challenge. Histories grow with the number of observations per user and can become large and unwieldy. Therefore, for computational tractability, we replace the complete history of observations with a representative summary. This representative summary, denoted by $Z \in \mathbb{R}^d$, is a vector of sufficiently large dimension $d$. This dimension $d$—determined by a hold-out validation set—is a constant and does not grow with the number of observations. The summary is computed as a function of history, $Z_n = f(H_n)$, where $f$ satisfies the Markov property, i.e., any two histories $h_1$ and $h_2$ mapped by $f$ to the same summary $z$, also have the same probabilities for their next observation. In other words, the summary is as predictive of future observations as the complete history of observations, and therefore summarizes all the information in the history.

Predictive state representations or PSRs (Littman et al. 2001, Thon and Jaeger 2015) provide a way to compute such representative summaries of the history of observations. 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 and an update function. The update function is used to update the internal state with every input of the sequence. RNNs are trained to minimize an output prediction error based on the sequence of inputs. Thus, the internal state is trained to “remember” the relevant information from the sequence.

We build an RNN that takes as input the sequence of tuples \((O_n, B_n)\) and predicts the subsequent observation \(\hat{O}_{n+1}\). The RNN minimizes the mean squared loss between the predicted and realized observations. The internal state \(Z_n\) is the PSR of the history of observations until auction \(n\). The architecture of the RNN and the dimension of the internal state are determined by a hold-out validation set as described in Web Appendix A.

To summarize, the estimation method proceeds in two steps. First, we train the RNN to generate PSRs at each auction. Next, we implement the semi-gradient TD method described in the previous subsection by replacing the complete history of observations with the PSR computed in the first step. Note that the pre-processing step of computing PSRs is only introduced for computational tractability.

4. **Empirical Analysis and Validation**

In this section, we demonstrate how the proposed method can be applied to measure the incremental impact of a campaign on conversions, as well as to quantify the impact of individual impressions. Furthermore, we validate the estimate of the impact of the campaign using a field experiment.
4.1. 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.

4.2. 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.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>List of Verticals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts &amp; Entertainment</td>
<td>Jobs &amp; Education</td>
</tr>
<tr>
<td>Autos &amp; Vehicles</td>
<td>Law &amp; Government</td>
</tr>
<tr>
<td>Beauty &amp; Fitness</td>
<td>News</td>
</tr>
<tr>
<td>Books &amp; Literature</td>
<td>Online Communities</td>
</tr>
<tr>
<td>Business &amp; Industrial</td>
<td>People &amp; Society</td>
</tr>
<tr>
<td>Computers &amp; Electronics</td>
<td>Pets &amp; Animals</td>
</tr>
<tr>
<td>Finance</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Food &amp; Drink</td>
<td>Reference</td>
</tr>
<tr>
<td>Games</td>
<td>Science</td>
</tr>
<tr>
<td>Health</td>
<td>Shopping</td>
</tr>
<tr>
<td>Hobbies &amp; Leisure</td>
<td>Sports</td>
</tr>
<tr>
<td>Home &amp; Garden</td>
<td>Travel &amp; Transportation</td>
</tr>
<tr>
<td>Internet &amp; Telecom</td>
<td>World Localities</td>
</tr>
</tbody>
</table>
For each auction, the targeting algorithm computes a bid using the history of (1) vertical information obtained from prior auctions, (2) impressions (total number and timing) served to the user, and (3) conversions from the user. In our data, we observe all the inputs to the targeting algorithm. Thus, our data satisfy the unconfoundedness assumption discussed in Section 3.2.

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.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Summary of Campaign Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>Users</td>
<td>82,612</td>
</tr>
<tr>
<td>Auctions</td>
<td>200,498,908</td>
</tr>
<tr>
<td>Impressions</td>
<td>400,239</td>
</tr>
<tr>
<td>Conversions</td>
<td>1,101</td>
</tr>
</tbody>
</table>

4.3. Impact of the Campaign

4.3.1. RCT. 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 intent-to-treat (ITT) 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 \(\frac{219}{35632} = .61\%\).

\(^1\) We performed a variety of randomization checks and found no evidence against proper randomization.
respectively. Thus, the ITT effect of the campaign is .72%. The percentage lift in conversions attributable to the campaign is therefore $\frac{1.33 - 0.67}{0.61} = 116.8\%$. The 95% bootstrapped confidence interval\(^2\) for the lift is [89.93\%, 145.68\%].

In the rest of this subsection, 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.

4.3.2. Commonly Used Observational Methods. We now compute the impact of the campaign using three classes of methods that are commonly used by researchers and practitioners: (1) regression adjustment—linear and logistic models, (2) propensity score matching, and (3) stratified regression adjustment. These methods compute the impact of the campaign by imposing different degrees of structure over the data generating process. Regression adjustment relies on an outcome model, while propensity score matching relies on an exposure model. Stratified regression uses an exposure model to group observations into strata and uses an outcome model within each stratum. For a rigorous treatment of the underlying assumptions and estimation procedures for these methods, see Gordon et al. (2019).

To account for the dynamic nature of RTB, we conduct our analyses at the user-day level\(^3\). Apart from impressions, the covariates used in each of the models include the average bid submitted by the advertiser, the total number of pages browsed by the user, and the number of pages browsed in each vertical. The average bid allows for controlling for targeting-induced endogeneity, while the page-view variables allow for reduction in activity bias (Hoban and Arora 2021).

\(^2\) All the confidence intervals in this section are computed by bootstrapping at the user level.

\(^3\) To check for robustness, we have also conducted the analyses at the user-hour and user-week levels, with no qualitative differences in the results.
It is worth noting that these methods estimate the average treatment effect on the treated (ATT). To enable comparison with the RCT estimate, we compute the ITT effect by multiplying the ATT with the fraction of units that are exposed in the treatment group (Angrist and Imbens 1995). The percentage lift is then computed as

\[
\text{lift} = \frac{\text{ITT}}{\eta_u - \text{ITT}} \times 100
\]

where \( \eta_u \) is the average number of conversions observed per unit.

Table 3 provides a summary of the lifts estimated by each of the methods. Comparing with the RCT estimate (116.8%), it is evident that these methods produce severely biased estimates. The estimates range from -49.74% to 41.85%. The best performing model estimates the lift to be 41.85%, underestimating the RCT lift by about 75%. Moreover, the best performing method and the set of covariates included in it could not have been determined a-priori, i.e., without the knowledge of the RCT estimate.

These results are consistent with the literature (Lewis et al. 2011, Gordon et al. 2019). Commonly used observational methods have been found to produce estimates with large bias, and the direction of the bias is unpredictable (Gordon et al. 2021). Gordon et al. (2019) hypothesize that the violation of the unconfoundedness assumption—i.e., researchers not observing all the inputs to the advertiser’s decision making—is the key driver of the bias. However, we show here that these methods produce biased estimates even when unconfoundedness is satisfied. The bias remains even when the unit of analysis is more granular than the aggregate user level. Thus, we conclude that these methods suffer from a large inductive bias—they are not sufficiently expressive to model the dynamic nature of ad serving in RTB, and further make strong assumptions on the data generating process.
Table 3 Summary of Lift Results - Commonly Used Observational Methods

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Regression Adjustment (Logistic)</th>
<th>Regression Adjustment (Linear)</th>
<th>Propensity Score Matching</th>
<th>Stratified Regression (Logistic)</th>
<th>Stratified Regression (Linear)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumPages</td>
<td>-1.45%</td>
<td>2.94%</td>
<td>2.88%</td>
<td>-12.82%</td>
<td>3.56%</td>
</tr>
<tr>
<td>AvgBid</td>
<td>39.20%</td>
<td>41.85%</td>
<td>32.29%</td>
<td>-2.33%</td>
<td>32.27%</td>
</tr>
<tr>
<td>NumPages + AvgBid</td>
<td>5.19%</td>
<td>9.51%</td>
<td>15.24%</td>
<td>-14.14%</td>
<td>17.43%</td>
</tr>
<tr>
<td>NumPages + AvgBid + VerticalCounts</td>
<td>5.14%</td>
<td>9.90%</td>
<td>15.56%</td>
<td>-49.74%</td>
<td>18.91%</td>
</tr>
</tbody>
</table>

4.3.3. 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 3.3 depends only on the sequence of advertiser’s observations \( \{O_1, O_2, \ldots\} \), bids \( \{B_1, B_2, \ldots\} \), and conversion indicators \( \{R_1, R_2, \ldots\} \).

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 \( n \).

To estimate the impact of the campaign, we compute the average conversion rates per auction \( \rho_{\pi_A} \) and \( \rho_{\pi_N} \) corresponding, respectively, to the advertiser’s policy \( \pi_A \) and the null policy \( \pi_N \). \( \rho_{\pi_A} \) 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 $\rho_{\pi_N} = 2.47 \times 10^{-6}$. The estimated lift is then $\frac{5.49 - 2.47}{2.47} = 122.19\%$. 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.

4.4. Impact of Individual Impressions

In addition to measuring the impact of the campaign, the MDP framework can be used to measure the individual impacts of the impressions served. As described in Section 3.2, this is achieved through the Q-function $Q_{\pi_N}$. For each impression, we compute $Q_{\pi_N}(H_n, B_n) - Q_{\pi_N}(H_n, 0)$, where $n$ is the auction through which the impression is served. This measures the incremental impact of the bid $B_n$ on the total conversions relative to a counterfactual bid of zero.

Figure 3 shows the density plot of the individual impacts of each of the 400,239 impressions served through the campaign. These impression impacts show a large heterogeneity, with estimates ranging from $1.5 \times 10^{-6}$ to $7.16 \times 10^{-6}$. The mean impression impact is $2.18 \times 10^{-6}$.

Exploring how an impression’s impact changes as a function of the user state allows advertisers to accurately value impression opportunities. It also allows advertisers and researchers to explore possible moderators of an impression’s impact. The effect of such
moderators can be computed directly from their effect on the Q-function. This can help contribute to the large body of literature exploring moderators such as an individual’s stage in the purchase funnel (Hoban and Bucklin 2015), frequency and timing of past impressions (Sahni 2015, Braun and Moe 2013), and content based targeting (Goldfarb and Tucker 2011). Furthermore, understanding the effect of moderators on ad response allows advertisers to examine the potential impact of a shift in targeting strategy. Thus, they play a critical role in optimizing campaign performance.

5. Conclusion

In this paper, we propose an observational framework to measure the incremental impact of display advertising campaigns, as well as the impact of individual impressions. 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.
We now list several avenues of future research and potential extensions. 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


Forbes (2015) As brands turn to digital advertising to reach the right audience, focus on validation is increasing. Website, URL https://tinyurl.com/yb8z14to, accessed 2021-02-03.


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


Johnson GA, Lewis RA, Nubbemeyer EI (2017) Ghost ads: Improving the economics of measuring online ad


grasping with deep learning and large-scale data collection. The International Journal of Robotics

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. The Quarterly

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.


learning with arbitrary smooth function approximation. NIPS, 1204–1212.


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, \ldots\} and bids \{B_1, \ldots\} to train a recurrent neural network (RNN) that generates a sequence of predictive state representations (PSRs) \{Z_1, \ldots\}. Next, we apply a temporal difference algorithm to the sequence of PSRs \{Z_1, \ldots\}, bids \{B_1, \ldots\}, and rewards \{R_1, \ldots\} 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 4. 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, \ldots, x_N\} to compute a sequence of “internal states” \{h_1, \ldots, 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

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

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

\[ p_n = \text{sigmoid}(W_p x_n + U_p h_{n-1}) \]
\[ q_n = \text{sigmoid}(W_q x_n + U_q h_{n-1}) \]

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 units 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 6 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 6 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.