Online Inference for Advertising Auctions*

Caio Waisman    Harikesh S. Nair    Carlos Carrion    Nan Xu

This draft: August 22, 2019

Abstract

Advertisers that engage in real-time bidding (RTB) to display their ads commonly have two goals: learning their optimal bidding policy and estimating the expected effect of exposing users to their ads. Typical strategies to accomplish one of these goals tend to ignore the other, creating an apparent tension between the two. This paper exploits the economic structure of the bid optimization problem faced by advertisers to show that these two objectives can actually be perfectly aligned. By framing the advertiser’s problem as a multi-armed bandit (MAB) problem, we propose a modified Thompson Sampling (TS) algorithm that concurrently learns the optimal bidding policy and estimates the expected effect of displaying the ad while minimizing economic losses from potential sub-optimal bidding. Simulations show that not only the proposed method successfully accomplishes the advertiser’s goals, but also does so at a much lower cost than more conventional experimentation policies aimed at performing causal inference.

Keywords: Causal inference, multi-armed bandits, advertising auctions

*The authors are part of JD Intelligent Ads Lab. The views represent that of the authors, and not JD.com. We thank Tong Geng, Jun Hao, Xiliang Lin, Lei Wu, Paul Yan, Bo Zhang, Liang Zhang and Lizhou Zheng for their collaboration; Günter Hitsch and Sanjog Misra for helpful comments. Please contact the authors at caio.waisman@kellogg.northwestern.edu (Waisman); harikesh.nair@stanford.edu (Nair); carlos.carrion@jd.com (Carrion); or nan.xu@jd.com (Xu) for correspondence.
1 Introduction

RTB systems, which leverage auctions to programmatically allocate user impressions to multiple competing advertisers, continue to enjoy widespread success in digital advertising (Varian, 2009; Yuan et al., 2013; Choi et al., 2017). Assessing the effectiveness of such advertising remains a lingering challenge in research and practice. This paper presents a new method to deliver causal inference on advertising bought through such mechanisms. Our method leverages the economic structure of second-price sealed-bid auctions, which are ubiquitous in RTB systems, embedded within a MAB setup for online adaptive experimentation. The algorithm we present develops estimates of the causal effects of advertising while minimizing the costs of experimentation to the advertiser by simultaneously learning the optimal bidding policy that maximizes her expected payoffs from auction participation. Our approach pertains to a nascent literature that leverages MABs for causal inference, more broadly.

There are two main challenges to developing an online experimental design for RTB ads. Firstly, measuring the average treatment effect (ATE) of advertising requires comparing outcomes of users who are exposed to ads with outcomes of users who are not. The complication is that ad-exposure is not under complete control of the experimenter because it is determined by an auction. This precludes simpler bandit designs for measuring the ATE in which exposure or non-exposure to ads are arms and outcomes under exposure or non-exposure are rewards. Instead, we need a design in which the experimenter controls only an input to the auction (bids), but wishes to perform inference on the effect of a stochastic outcome induced by this input (ad-exposure). In addition, in many advertising situations it is in the interest of the advertiser to account for observable heterogeneity when assessing ad-effectiveness. In such cases, inference would have to be performed on heterogeneous treatment effects, and the object of estimation would no longer be the ATE, but rather the conditional average treatment effect (CATE) at a series of values of observed variables.

A second complication is the need to balance the goal of learning the expected effect of ad-exposure (“inference goal”) with the goal of learning the optimal bidding policy for the advertiser (“economic goal”). The tension is that finding the best bidding policy does not guarantee estimating ad-effectiveness and vice-versa. At one extreme, with a bidding policy that delivers on the economic goal, the advertiser could win most of the time, making it difficult to measure ad-effectiveness since outcomes with no ad-exposures would
be scarcely observed. At the other extreme, with pure bid randomization the advertiser could estimate unconditional ad-effectiveness (the \textit{ATE}) or how ad-effectiveness varies with observed heterogeneity (the \textit{CATE}s) and deliver on the inference goal, but may end up incurring large economic losses in the process.

The contribution of this paper is to present a bandit design and statistical learning framework that address both these considerations. In our design, observed heterogeneity is summarized by a context, \( x \), bids form arms, and the advertiser’s economic payoffs from the bids form the rewards, so that the best arm maximizes the advertiser’s expected payoff from auction participation given \( x \). Exploiting the economic structure of second-price sealed-bid auctions, we show that, under conditions we outline, the bid that maximizes the advertiser’s expected payoff from auction participation at a given \( x \) corresponds to the \textit{CATE} of the ad at the value \( x \), or \( \text{ATE}(x) \). Therefore, in our experimental design, discovering the best bid and estimating ad effectiveness are the same task, so that the inference and economic goals are perfectly aligned.

To implement the design, we present a TS algorithm customized to our auction environment, trained online via a Markov Chain Monte Carlo (MCMC) method. The algorithm adaptively chooses bids across rounds based on current estimates of which arm is the optimal one. These estimates are updated on each round via MCMC through Gibbs sampling, which leverages data augmentation to impute the missing potential outcomes and the censored winning bids in each round of the bandit. Simulations show that the algorithm is able to recover treatment effect heterogeneity as represented by the \textit{CATE}s of advertising and considerably reduces the expected costs of experimentation compared to non-adaptive “A/B/n” tests.

\section{1.1 Related literature}

Our paper lies at the intersection of three broad fields of study: causal inference, multi-armed bandits and auction theory. We will now describe how it relates to the existing literature, confining our attention to studies that touch at least two of these three topics.

Performing causal inference with bandits is complicated by the adaptive nature of data collection, wherein future data collection depends on the data already collected. Although bandits possess attractive properties in finding the best arm, estimates of arm-specific expected rewards typically exhibit adaptive bias (Xu et al., 2013; Villar et al., 2015).
In particular, Nie et al. (2018) show that archetypal bandit algorithms such as Upper Confidence Bound (UCB) and TS compute estimates of arm-specific expected rewards that are biased downwards. Adaptive data collection also complicates frequentist statistical inference, requiring adjustments for computation of valid standard errors, test statistics and confidence intervals.

Existing approaches to inference with bandits differ based on whether they pertain to the \textit{offline} setting, where pre-collected data is available to the analyst, or the \textit{online} setting, where data arrive sequentially, with the online literature being relatively more recent. Unlike online methods, offline methods are meant to be implemented \textit{ex-post}, which implies that the data collection, though done sequentially, is typically not made explicitly to facilitate inference upon its completion. Also, the methods are meant for retrospective application on logged data, which means that data collection does not explicitly reflect in real-time the progress made towards the inferential goal. Offline methods for logged bandit data that reflect adaptive data collection include Strehl et al. (2010), Dudík et al. (2011), Li et al. (2015), Swaminathan and Joachims (2015), Jiang and Li (2016), Thomas and Brunskill (2016), Athey and Wager (2017), Wang et al. (2017b), Diemert et al. (2018), Deshpande et al. (2018), Kallus (2018a), Nie et al. (2018) and Sawant et al. (2018).

This paper relates more closely to the online stream. Online methods for finding the best arm while correcting for adaptive bias include Goldenshluger and Zeevi (2013), Bastani and Bayati (2015) and Nie et al. (2018), who suggest data-splitting by forced-sampling of arms at prescribed times, and Dimakopoulou et al. (2018), who correct for the bias via balancing and inverse probability weighting. Online methods to perform frequentist statistical inference that is valid for bandits, but which avoid issues of explicitly bias-correcting estimates of arm-specific expected rewards, are presented in Yang et al. (2017), Jamieson and Jain (2018) and Ju et al. (2019).

Nevertheless, this paper has a different focus on inference compared to the previous literature. Broadly speaking, the above methods aim to either find the best arm or learn without bias the expected reward associated with the best arm. In contrast, our goal is to obtain an unbiased estimate of the effect of an action (exposure to advertising) that is imperfectly obtained by pulling arms (bids). Therefore, in our setup, the target treatment whose effect is to be learned is not an arm, but a shared stochastic outcome that arises from pulling arms. Hence, arms are more appropriately thought of as \textit{instruments} for treatments, which makes our setup the online analogue of an offline encouragement design from the program evaluation literature (see, for example, Imbens and Rubin, 1997).
Our setup also shares similarities with the IV-bandit setup of Kallus (2018b), in which there is a difference between the treatment-arm pulled and the treatment-arm applied due to the possibility of user non-compliance. However, the difference between the pulled and applied treatments, which is important to the design here, is not a feature of the design considered by Kallus, because the pulled and applied treatments belong to the same set in his design. Also, unlike Kallus’ setup, where exposure to a treatment is the outcome of a choice by the user to comply with the pulled arm, exposure here is obtained via a multi-agent game that is not directly affected by the user (auction), thus characterizing a different exposure mechanism.

Bandits have been embedded explicitly within the structural causal framework of Pearl (2009) in a series of recent papers by Bareinboim et al. (2015), Lattimore et al. (2016) and Forney et al. (2017). Our paper is related to this stream as our application is a specific instance of a structural causal model tailored to the auction setting: we assume the existence of a probabilistic generative process that is the common shared causal structure behind the distributions of the rewards for each arm. As this stream has emphasized (for a computational advertising example, see Bottou et al., 2013), the link to the model in our application is helpful to making progress on the inference problem. This approach has also been followed by other papers in economics (see, for example, the references in Bergemann and Välimäki, 2008) and marketing (Misra et al., 2019) that study pricing problems where firms aim to learn the optimal price from a grid of prices, corresponding to arms, which share the same underlying demand function.

Finally, our approach is grounded in auction theory and also relates to bandits as applied to RTB for digital advertising. In general, bandits and more general reinforcement learning approaches have been used to learn optimal bidding policies (Cai et al., 2017; Wang et al., 2017a; Wu et al., 2018; Jin et al., 2018), but not to estimate treatment effects. We depart from these approaches in that our goal is to also perform causal inference. Further, we achieve this goal by leveraging key properties of the auction format, thus contributing to a nascent literature, to our knowledge, on direct applications of auction theory to enable causal inference. While several studies combined experimentation with auction theory, their goals were to identify optimal policies such as bids as in the aforementioned studies, reserve prices (Cesa-Bianchi et al., 2014; Austin et al., 2016; Ostrovsky and Schwarz, 2016; Pouget-Abadie et al., 2018; Rhuggenaath et al., 2019) or auction formats (Chawla et al., 2016), not to estimate the causal effect of an action determined by the auction.
2 The ad display auction model

A risk neutral advertiser (she) participates in a second-price sealed-bid auction to display her ad to a consumer (he). Her total expected payoff from the auction, in monetary units, is given by:

\[
\tilde{\pi}(b|x) = \Pr \{ B_{CP} \leq b|x \} \times \mathbb{E} \left[ Y(1) - B_{CP} \mid B_{CP} \leq b; x \right] \\
+ \Pr \{ B_{CP} > b|x \} \times \mathbb{E} \left[ Y(0) \mid B_{CP} > b; x \right],
\]

where \( Y(1) \) is the payoff the advertiser receives if she wins the auction, \( Y(0) \) is the payoff she receives if she loses, \( B_{CP} \) is the highest bid against which she competes, \( b \) is the bid she submits, and \( x \) is a variable that characterizes the auction and which can take \( p \) different values, so that \( x \in X \equiv \{ x_1, ..., x_p \} \).\(^1\) The expectations in (1) are taken with respect not only to \( B_{CP} \), as in standard auction models, but also to \( Y(1) \) and \( Y(0) \), and the advertiser’s optimization problem is to maximize (1) with respect to \( b \). Thus, we assume that the advertiser faces no inter-auction restrictions such as a budget constraint. To ensure that the solution to this optimization problem is well-defined we maintain the following common technical assumption regarding the joint distribution of \{\( Y(1), Y(0), B_{CP} \)\} conditional on \( x \), which we denote by \( F(\cdot, \cdot, \cdot | x) \).

**Assumption 1.**

(i) The joint distribution \( F(\cdot, \cdot, \cdot | x) \) admits a continuous density, \( f(\cdot, \cdot, \cdot | x) \), over \( \mathbb{R}_+^3 \) for all \( x \).

(ii) The density of \( B_{CP} \) given \( x \), \( f_{CP}(\cdot | x) \), is strictly positive in the interior of \( \mathbb{R}_+ \) for all \( x \).

(iii) \( \mathbb{E} [Y(1) | x] < \infty \), \( \mathbb{E} [Y(0) | x] < \infty \), and \( \mathbb{E} [B_{CP} | x] < \infty \) for all \( x \).

In standard auction models the term \( Y(0) \) is set to zero. However, this convention is not fitting to our setting given how we interpret the payoff terms \( Y(1) \) and \( Y(0) \). In particular, a consumer might have a baseline propensity to purchase the advertiser’s product even if he is not exposed to her ad, which is associated with the term \( Y(0) \). Exposure to the ad might affect this propensity, which implies that \( Y(1) \neq Y(0) \).

\(^1\)In our MAB setup, \( x \) is the context to which the auction belongs. It can be obtained, for example, from a vector \( Z \) of observable display opportunity variables that can include, for example, user and publisher characteristics, with \( p \) being the total number of different combinations of values across all elements of \( Z \). This procedure is analogous to the segmentation of consumers in Misra et al. (2019).
3 The advertiser’s dual objective

Standard auction theory models assume that the joint distribution $F(\cdot, \cdot, \cdot | \cdot)$ is known, in which case computing $b^*(\cdot) \equiv \arg\max_b \pi(b | \cdot)$ is straightforward, as is measuring the heterogeneous expected effects of the ad, which is given by $ATE(\cdot) \equiv E[Y(1) - Y(0) | \cdot]$. However, in reality the advertiser has to collect data with information on $F(\cdot, \cdot, \cdot | \cdot)$ to estimate $b^*(\cdot)$ and $ATE(\cdot)$. Each observation $i$ in these data is an ad display auction, and for each $i$ the advertiser observes a vector $\{Y_i, D_i, B_{CP,i}, b_i, x_i\}$, where $D_i \equiv 1 \{B_{CP,i} \leq b_i\}$, $Y_i \equiv D_i \times Y_i(1) + (1 - D_i) \times Y_i(0)$ and $B_{CP,i} \equiv \min \{B_{CP,i}, b_i\}$. The structure of these data is given in Table 1 below.

| $i$ | $b_i$ | $x_i$ | $D_i$ | $Y_i$ | $Y_i(1)$ | $Y_i(0)$ | $B_{CP,i}$ |
|-----|------|------|------|------|---------|---------|-----------|
| 1   | $b_1$ | $x_1$ | 1    | $y_1$ | $y_1$   | $-       | $b_{CP,1}$|
| 2   | $b_2$ | $x_2$ | 0    | $y_2$ | $-       | $y_2$   | $b_2$    |
| ... | ...  | ...  | ... | ...  | ...     | ...     | ...       |

Notice that these data suffer from two issues. The first is the “fundamental problem of causal inference” (Holland, 1986): $Y(1)$ and $Y(0)$ are never observed at the same time. The second is a censoring problem: $B_{CP}$ is only observed when the advertiser wins the auction; otherwise, all she knows is that it was larger than the bid she submitted. Hence, the observed data have a similar structure to the one in the model defined by Amemyia (1984) as the Type 4 Tobit model.

Before proceeding to analyze the advertiser’s learning goals, we state the following assumption, which we maintain throughout the analysis.

**Assumption 2.**

\{Y_i(1), Y_i(0), B_{CP,i}\} \overset{iid}{\sim} F(\cdot, \cdot, \cdot | x_i) \text{ and } x_i \overset{iid}{\sim} F_x(\cdot).

Assumption 2 imposes restrictions on the data generating process (DGP), which is further illustrated in Figure 1. For instance, if the same consumer appeared more than once and if his potential outcomes $Y(1)$ and $Y(0)$ were serially correlated, this condition would not hold. In turn, if competing bidders solved a dynamic problem because of
longer-term dependencies, budget or impression constraints, $B_{CP}$ could become serially correlated as a result, in which case this condition would also be violated.

![Figure 1: DGP](image)

### 3.1 Economic goal

The advertiser’s economic goal is to learn $b^*(\cdot)$ from a sequence of auctions. A policy that aims to achieve this goal should quickly learn $b^*(\cdot)$ while minimizing losses from occasional sub-optimal bidding. Given Assumptions 1 and 2, $b^*(x)$ is well-defined and common across auctions for every $x$, which implies that the advertiser’s economic goal lends itself naturally to be represented as a contextual MAB problem.

In such setting, the advertiser considers a set of $r_x = 1, \ldots, R_x$ arms (so that we allow the grid of arms to change across different contexts), each of which associated with a bid, $b_{rx}$. The advertiser’s goal can be expressed as minimizing the expected regret from potentially bidding sub-optimally over a sequence of auctions while learning $b^*(\cdot)$. Hence, we implicitly assume that for each $x$ the grid contains the optimal bid, $b^*(x)$. Algorithms used to solve MAB problems typically base the decision of which bid to play in round $t$, $b_t$, on a tradeoff between randomly picking a bid to obtain more information about its associated payoff (exploration) and the information gathered until then on the optimality of each bid.
(exploitation). The existing information at the beginning of round \( t \) is a function of all data collected until then, which we denote by \( I_{t-1} \). Stacking the data presented in Table 1 across auctions for each round \( \tau \), we write \( I_{t-1} = \{b_\tau, x_\tau, D_\tau, Y_\tau, B_{CP,\tau}, \omega_\tau\}^{\tau-1}_{\tau=1} \). The \( \omega \)s are seeds, independent from all other variables, required for randomization depending on which algorithm is used.

Notice two points of departure between this setup and the usual MAB problem. First, in the latter, each arm is associated with a different DGP, so it is commonly assumed that the reward draws are independent across arms. In our setting, this is not true: given the economic structure of the problem, conditional on \( x \) the values of \( \{Y(1), Y(0), B_{CP}\} \) are the same regardless of which arm is pulled, which creates correlation between rewards across arms. Second, on pulling an arm we observe three different forms of feedback: an indicator for whether we obtain treatment (ad-exposure), the highest competing bid conditional on treatment, and the reward. This contrasts with the canonical case in which the reward forms the only source of feedback.

### 3.2 Inference goal

The advertiser’s inference goal is to estimate \( ATE(\cdot) \) from data. The main challenge in achieving this goal is the missing data problem. A well-known solution to it would be for the advertiser to randomize treatment, that is, ad-exposure. However, in this setting this is not feasible since the treatment is determined by the outcome of the auction, which is not under the advertiser’s complete control because she does not determine the highest competing bid.

Nevertheless, the advertiser does have full control over her bids. Consequently, an identification strategy to recover \( ATE(\cdot) \) from data could make use of bid randomization.\(^2\) Notice that this generally is not sufficient to yield identification of \( ATE(\cdot) \) if the relationship between \( \{Y(1), Y(0)\} \) and \( B_{CP} \) remains unrestricted. However, under the following assumption bid randomization becomes equivalent to random ad-exposure, and therefore yields identification of \( ATE(\cdot) \). We discuss this condition in more detail in the next section.

**Assumption 3.**
\[ \{Y(1), Y(0)\} \perp\!\!\!\!\!\!\perp B_{CP} | x. \]

\(^2\)This has also been noted by Lewis and Wong (2018).
4 Aligning the economic and inference goals

The two aforementioned goals are related as they both rely on learning the joint distribution \( F(\cdot, \cdot, \cdot | \cdot) \) from data. Nevertheless, common solutions to achieve one of these goals often ignore the other one, creating a misalignment between the two.

First, consider the use of an algorithm to solve the MAB problem to learn \( b^*(\cdot) \) and assume that \( b^*(\cdot) \) is such that the advertiser would almost always win by bidding optimally. An efficient algorithm would quickly converge, eventually providing few auctions in which \( D = 0 \), yielding few observations of \( Y(0) \), and making conventional estimates of \( ATE(\cdot) \) based on the data gathered via such algorithm imprecise.

Now assume that Assumption 3 holds and that the advertiser fully randomizes bids to recover \( ATE(\cdot) \). While many algorithms used to solve contextual MAB problems involve bid randomization, the extent to which this is done is determined by an explicit objective of minimizing losses from sub-optimal decisions. Full bid randomization is likely to yield considerably larger economic losses to the advertiser, which in many cases may preclude its use altogether.

Despite seeming irreconcilable, we now demonstrate that, under Assumptions 1 and 3, the advertiser’s goals are actually perfectly aligned. We express this in the following proposition.

**Proposition 1.** If Assumptions 1 and 3 hold, \( b^*(x) = \max \{0, ATE(x)\} \).

**Proof.** To prove Proposition 1, we first rewrite equation (1):

\[
\bar{\pi}(b|x) = \Pr \{B_{CP} \leq b|x\} \mathbb{E}[Y(1) - B_{CP}|B_{CP} \leq b;x] + \Pr \{B_{CP} > b|x\} \mathbb{E}[Y(0)|B_{CP} > b;x] \\
= \Pr \{B_{CP} \leq b|x\} \{\mathbb{E}[Y(1) - Y(0)|B_{CP} \leq b;x] - \mathbb{E}[B_{CP}|B_{CP} \leq b;x]\} + \mathbb{E}[Y(0)|x] \\
= \Pr \{B_{CP} \leq b|x\} \{ATE(x) - \mathbb{E}[B_{CP}|B_{CP} \leq b;x]\} + \mathbb{E}[Y(0)|x] \\
= \Pr \{B_{CP} \leq b|x\} \{ATE(x) - \mathbb{E}[B_{CP}|B_{CP} \leq b;x]\} + \mathbb{E}[Y(0)|x] \\
= \int_0^b [ATE(x) - u] f_{CP}(u|x) du + \mathbb{E}[Y(0)|x], \tag{2}
\]

where the third equality follows from Assumption 3. Ignoring the term \( \mathbb{E}[Y(0)|x] \) in (2), which is without loss since it does not depend on \( b \) and thus has no impact on the optimization problem, notice that \( \bar{\pi}(b|x) \) becomes a bidder’s expected payoff from...
a second-price sealed-bid auction in which this bidder’s private value equals $ATE(x)$. Because the advertiser cannot submit negative bids, when $ATE(x) \leq 0$ the optimal bid is $b(x) = 0$ since the integrand is negative. Otherwise, notice that the integrand is non-negative as long as $b \leq ATE(x)$, which implies that the optimal bid cannot be greater than $ATE(x)$. Because of Assumption 1, the density $f_{CP}(\cdot|x)$ is strictly positive, so that $b^*(x) = ATE(x)$.

Proposition 1 is powerful because it implies that whenever displaying the auction is beneficial, that is, when $ATE(x) \geq 0$, the advertiser’s economic and inference goals are perfectly aligned, as learning $b^*(x)$ and estimating $ATE(x)$ consist of the same task. The usefulness of this result is that we obtain the object we would like to estimate and perform inference on, $ATE(x)$, as the identity of the best arm (that is, the best bid), rather than the expected value of its reward. Since typical MAB algorithms recover the identify of the best arm without bias, we are able to leverage them for inference on ad-effects without bias in an online environment. In turn, when ad-exposure is detrimental ($ATE(x) < 0$) the proposition converts this qualitative result into a clear economic policy as the advertiser would have no interest in displaying the ad in the first place, which can be guaranteed by a bid of zero. This result is obtained by exploiting the economic structure of the advertiser’s problem, so the conditions for it to be true warrant further comments.

Assumption 3 is key. From an auction theory perspective, it is intuitive since it bears a private values content: it implies that, conditional on $x$, knowledge of $B_{CP}$ has no effect on the bidder’s assessment of $\{Y(1), Y(0)\}$, and, consequently, on her assessment of her willingness-to-pay (valuation). Assumption 2 is not required to establish Proposition 1, but it does justify framing the seller’s optimization problem as a contextual MAB. If instead $Y(1), Y(0)$ or $B_{CP}$ were serially correlated across auctions the advertiser’s dynamic problem would fit into a more general type of reinforcement learning problem. This would also be the case if the advertiser faced inter-auction restrictions such as a budget constraint, which we have ruled out by assumption. In turn, Assumption 1 not only is mild but also relatively common in auction models and is made solely for tractability. The requirement that $f_{CP}(\cdot|x)$ is strictly positive in the interior of $\mathbb{R}_+$ guarantees that the optimal bid is unique, but is actually stronger than what is required to ensure uniqueness. A sufficient condition is that $f_{CP}(\cdot|x)$ is strictly positive on a neighborhood around $ATE(x)$. Finally, it is important to note that Proposition 1 generally holds for strategy-proof mechanisms, of which the second-price sealed-bid auction is just one example.\footnote{Other examples are ascending auctions, general Vickrey-Clarke-Groves (VCG) mechanisms and...}
5  Bidding Thompson Sampling (BTS) algorithm

We now propose a procedure to achieve the advertiser’s goals, which is a modified TS algorithm. Since it aims to learn the advertiser’s optimal bid, we will refer to it as Bidding Thompson Sampling (BTS).

5.1  General procedure

The TS algorithm (Thompson, 1933) is a Bayesian heuristic to solve MAB problems. TS typically starts by parametrizing the distribution of rewards associated with each arm. Since our problem departs from standard MAB problems in that the DGP behind each of the arms — that is, the distribution $F(\cdot, \cdot, \cdot | \cdot)$ — is the same, we choose to parametrize it instead and denote our vector of parameters of interest by $\theta$. Expected rewards depend on $\theta$, so we will often write $\bar{\pi}(b|x, \theta)$.

The algorithm runs while a criterion, $c_t$, is below a threshold, $T$. After round $t$, the prior over $\theta$ is updated by the likelihood of all data gathered by the end of round $t$, $I_t$. We denote the number of observations gathered on round $t$ by $n_t$ and the total number of observations gathered by the end of round $t$ by $N_t = \sum_{\tau=1}^t n_\tau$. If $n_t = 1$ for all $t$ the algorithm proceeds auction by auction. We present it in this way to accommodate batch updates. Given the posterior distribution of $\theta$ given $I_t$, we calculate

$$\psi_t(b_{r_x}|x) \equiv \Pr(\text{arm } r_x \text{ is the best arm } | I_t; x) \quad (3)$$

and update the criterion $c_t$. If the algorithm continues, on round $t + 1$ arm $r_x$ is pulled for each observation with context $x$ with probability $\psi_t(b_{r_x}|x)$; otherwise, the arm with highest probability of being the optimal one, $b^*_t(x) \equiv \arg \max_b \psi_t(b|x)$, is identified as

---

4See Scott (2010) for an application to computational advertising and Russo et al. (2018) for an overview.
such. The generic structure of the TS algorithm is outlined below.

| **Algorithm 1: Thompson Sampling** |
|-----------------------------------|
| 1 Set $p(\theta), \psi_0(\cdot|\cdot), c_0$ and $T$. |
| **while** ($c_t < T$) **do** |
| 2 Pull arms according to $\psi_{t-1}(\cdot|\cdot)$. |
| 3 Combine new data with previously obtained data in $I_t$. |
| 4 Update the posterior distribution of $\theta$ with $I_t$. |
| 5 Compute $\psi_t(\cdot|\cdot), c_t$ and $b^*_t(\cdot)$. |
| **end** |
| 6 Identify $b^*_T(\cdot)$ as optimal arm. |

### 5.2 Parametrizing distribution of rewards

We now present the specific parametrization we use in our problem. Let $X_i$ be the following $p-$dimensional vector of mutually exclusive dummies:

$$X_i \equiv \begin{bmatrix} \mathbb{1}\{x_i = x_1\}, \mathbb{1}\{x_i = x_2\}, ..., \mathbb{1}\{x_i = x_p\} \end{bmatrix}'. \tag{4}$$

Notice that there is a one-to-one correspondence between the vector $X_i$ and the variable $x$. Hence, we will use them interchangeably whenever it does not cause confusion. We assume that

$$\begin{bmatrix} \log Y_i(1) \\ \log Y_i(0) \\ \log B_{CP,i} \end{bmatrix} \overset{iid}{\sim} N \left( \begin{bmatrix} X_i' \delta_1 \\ X_i' \delta_0 \\ X_i' \delta_{CP} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_0 & 0 \\ \rho \sigma_1 \sigma_0 & \sigma_0^2 & 0 \\ 0 & 0 & \sigma_{CP}^2 \end{bmatrix} \right) \equiv N \left( \begin{bmatrix} \Delta'X_i \\ X_i' \delta_{CP} \end{bmatrix}, \begin{bmatrix} \Sigma & 0 \\ 0' & \sigma_{CP}^2 \end{bmatrix} \right), \tag{5}$$

where $\Delta \equiv [\delta_1, \delta_0]$, and collect the parameters in $\theta \equiv [\delta'_1, \delta'_0, \delta'_{CP}, \sigma_1^2, \sigma_0^2, \sigma_{CP}^2, \rho]'$.

Notice that this parametrization directly imposes Assumption 3 and that since both potential outcomes are never observed simultaneously, $\rho$ is not point identified without further restrictions.\(^5\) Also notice that (5) implies that $ATE(X_i) = \exp \{X_i' \delta_1 + 0.5 \sigma_1^2\} - \exp \{X_i' \delta_0 + 0.5 \sigma_0^2\}$. Hence, since our interest is in $ATE(\cdot)$ and since it does not depend

\(^5\)However, it is possible to exploit the positive semidefiniteness of $\Sigma$ to partially identify $\rho$. See, for example, Vijverberg (1993) and Koop and Poirier (1997).
on $\rho$, we explicitly assume that $\rho = 0$. This assumption has the benefit of significantly simplifying the algorithm we present. A more general version that allows for $\rho \neq 0$ is given in the Appendix. Furthermore, notice that (5) also implies that:

$$
\pi(b|X_t, \theta) = \Phi \left( \frac{\log b - X'_t\delta_{CP}}{\sigma_{CP}} \right) \times ATE(X_t)
- \Phi \left( \frac{\log b - X'_t\delta_{CP}}{\sigma_{CP}} - \sigma_{CP} \right) \times \exp \left\{ X'_t\delta_{CP} + 0.5\sigma_{CP}^2 \right\},
$$

(6)

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and we omit the terms that do not depend on $b$ for brevity. Finally, we do not allow the matrix $\Sigma$ to depend directly on $X_t$ for simplicity. Nothing in our procedure precludes such parametrization should the researcher prefer it.

### 5.3 Setting up priors using historical data

We choose independent normal-gamma priors, which are conjugate to the normal likelihood induced by the DGP in (5). For $k \in \{1, 0, CP\}$, we set:

$$
\sigma_k^{-2} \sim \Gamma(\alpha_k, \beta_k)
\delta_k|\sigma_k^2 \sim N\left(\mu_{\delta_k}, \sigma_k^2 A_k^{-1}\right),
$$

(7)

where $\{\alpha_k, \beta_k\}_{k \in \{1, 0, CP\}}$ are non-negative scalars, $\{\mu_{\delta_k}\}_{k \in \{1, 0, CP\}}$ are $p$-dimensional vectors and $\{A_k\}_{k \in \{1, 0, CP\}}$ are $p$-by-$p$ matrices. For the gamma distribution, the parametrization is such that if $G \sim \Gamma(\alpha, \beta)$, then $E[G] = \alpha/\beta$.

### 5.4 Drawing from posterior: Gibbs sampling

Implementing the algorithm requires computing updated probabilities, $\psi_t(\cdot|\cdot)$, which cannot be done analytically because of the missingness and censoring in the feedback data. Nevertheless, it is still possible to exploit conditional conjugacy via data augmentation and use the Gibbs sampling algorithm to obtain draws from the posterior distribution of $\theta$ given $I_t$. Using these draws we can then estimate $\psi_t(\cdot|\cdot)$ via Monte Carlo integration. We now describe the steps of this estimation procedure, which combines the methods introduced by Chib (1992) and Koop and Poirier (1997) in a single Gibbs sampling algorithm.
with data augmentation.

### 5.4.1 Data augmentation

The first step in our procedure is to draw the missing values from our data conditional on \((I_t, \theta)\). We begin by drawing the missing values \(\{\log B_{CP,i}\}_{i:D_i=0}\). Given \((I_t, \theta)\) and under (5), it follows that:

\[
\log B_{CP,i}^{\text{miss}} \bigg| D_i = 0, \log Y_i, \log B_{CP,i}, \log b_i, X_i, \theta \overset{d}{=} \log Y_i^{\text{miss}}(0) \bigg| D_i = 0, \log b_i, X_i, \delta_{CP}, \sigma_{CP}^2 \sim TN \left( X_i' \delta_{CP}, \sigma_{CP}^2, \log b_i, +\infty \right),
\]

where \(\overset{d}{=}\) means equality in distribution and \(TN (\delta_*, \sigma_*^2, l, u)\) denotes the truncated normal distribution with mean \(\delta_*\), variance \(\sigma_*^2\), lower truncation at \(l\) and upper truncation at \(u\).

Now we proceed to draw the missing values \(\{\log Y_i(1)\}_{i:D_i=0}\) and \(\{\log Y_i(0)\}_{i:D_i=1}\). Given \((I_t, \theta)\) and under (5), it follows that:

\[
\log Y_i^{\text{miss}}(1) \bigg| D_i = 0, \log Y_i, \log B_{CP,i}, \log b_i, X_i, \theta \overset{d}{=} \log Y_i^{\text{miss}}(1) \bigg| D_i = 0, X_i, \delta_1, \sigma_1^2 \sim N \left( X_i' \delta_1, \sigma_1^2 \right)
\]

and

\[
\log Y_i^{\text{miss}}(0) \bigg| D_i = 1, \log Y_i, \log B_{CP,i}, \log b_i, X_i, \theta \overset{d}{=} \log Y_i^{\text{miss}}(0) \bigg| D_i = 1, X_i, \delta_0, \sigma_0^2 \sim N \left( X_i' \delta_0, \sigma_0^2 \right).
\]

Defining

\[
\delta_i^{\text{miss}} = D_i X_i' \delta_0 + (1 - D_i) X_i' \delta_1
\]

and

\[
\sigma_i^{2,\text{miss}} = D_i \sigma_0^2 + (1 - D_i) \sigma_1^2,
\]

15
we can combine (9) and (10) into:

\[
\log Y_i^{\text{miss}} | \log Y_i, D_i, \log B_{\text{CP},i}, \log b_i, X_i, \theta \overset{d}{\sim} N \left( \theta_i^{\text{miss}}, \sigma_i^{2,\text{miss}} \right). \tag{13}
\]

### 5.4.2 Creating the “complete” data

Given a draw from the distributions given above, \( \{ \log Y_i^{\text{miss}}, \log B_{\text{CP},i}^{\text{miss}} \} \), we can construct the “complete” data:

\[
\begin{align*}
\log \bar{Y}_i(1) &= D_i \log Y_i + (1 - D_i) \log Y_i^{\text{miss}} \\
\log \bar{Y}_i(0) &= D_i \log Y_i^{\text{miss}} + (1 - D_i) \log Y_i \\
& \quad \quad \log \bar{B}_{\text{CP},i} = D_i \log \bar{B}_{\text{CP},i} + (1 - D_i) \log B_{\text{CP},i}^{\text{miss}}.
\end{align*} \tag{14}
\]

### 5.4.3 Drawing from posterior distribution

The last step is to draw new parameters from their full conditional distributions. Collect the parameters of the priors in \( \theta_{\text{prior}} \equiv \{ \mu_{\delta_k}, A_k, \alpha_k, \beta_k \}_{k \in \{1,0,\text{CP}\}} \). For ease of notation, we stack all the “complete” data by the end of round \( t \) in the following \( N_t \)-by-1 vectors: \( \log \bar{Y}_i(1), \log \bar{Y}_i(0), \log \bar{B}_{\text{CP},i}, D_i \) and \( \log b_i \). We also use the \( N_t \)-by-\( p \) matrix \( X_t \), whose \( i \)-th row is the vector \( X'_i \), and collect them all in \( I_t \equiv [\log \bar{Y}_i(1), \log \bar{Y}_i(0), \log \bar{B}_{\text{CP},i} \log b_i, D_t, X_t] \).

Finally, let \( \theta^{(q-1)} = [\theta'^{(q-1)}_1, \theta'^{(q-1)}_0, \theta'^{(q-1)}_{\text{CP}}, \sigma'^{(q-1)}_{\text{CP},1}, \sigma'^{(q-1)}_{\text{CP},0}, \sigma'^{(q-1)}_0] \) be the \( (q-1) \)-th draw of the vector of parameters. Given the structure of the model, it then follows that the full conditional distributions of the parameters simplify in the following way:

\[
\begin{align*}
\sigma^2_{\text{CP}} | \theta^{(q-1)}, \theta_{\text{prior}}, I_t & \overset{d}{\sim} \sigma^2_{\text{CP}} | \log \bar{B}_{\text{CP},t}, X_t, \mu_{\delta_{\text{CP}}}, A_{\text{CP}}, \alpha_{\text{CP}}, \beta_{\text{CP}} \\
\sigma^2_1 | \theta^{(q-1)}, \theta_{\text{prior}}, I_t & \overset{d}{\sim} \sigma^2_1 | \log \bar{Y}_t(1), X_t, \mu_{\delta_1}, A_1, \alpha_1, \beta_1 \\
\sigma^2_0 | \theta^{(q-1)}, \theta_{\text{prior}}, I_t & \overset{d}{\sim} \sigma^2_0 | \log \bar{Y}_t(0), X_t, \mu_{\delta_0}, A_0, \alpha_0, \beta_0
\end{align*} \tag{15}
\]
and, letting \( \sigma^2(q) \equiv \left[ \sigma^2_1(q), \sigma^2_0(q), \sigma^2_{CP}(q) \right]' \),

\[
\delta_{CP} \left| \begin{array}{c}
\sigma^2(q), \delta^{(q-1)}_{0}, \delta^{(q-1)}_{CP}, \theta_{prior}, I_t = d \end{array} \right| \sigma^2_{CP} \ \log \tilde{B}_{CP,t}, X_t, \mu_{\delta_{CP}}, A_{CP}
\]

\[
\delta^{(q)}_1 \left| \begin{array}{c}
\sigma^2(q), \delta^{(q-1)}_{0}, \delta^{(q-1)}_{CP}, \theta_{prior}, I_t = d \end{array} \right| \sigma^2_{1} \ \log \tilde{Y}_t(1), X_t, \mu_{\delta_1}, A_1
\]

\[
\delta^{(q)}_0 \left| \begin{array}{c}
\sigma^2(q), \delta^{(q-1)}_{0}, \delta^{(q-1)}_{CP}, \theta_{prior}, I_t = d \end{array} \right| \sigma^2_{0} \ \log \tilde{Y}_t(0), X_t, \mu_{\delta_0}, A_0.
\]

For completeness, because of the parametric forms in (5) and (7), we have that:

\[
\sigma_{CP}^{-2(q)} \left| \begin{array}{c}
\log \tilde{B}_{CP,t}, X_t, \mu_{\delta_{CP}}, A_{CP}, \alpha_{CP}, \beta_{CP}
\end{array} \right| \sim \Gamma \left( \alpha_{CP} + \frac{N_t}{2}, \beta_{CP} + \frac{1}{2} \left[ (\log \tilde{B}_{CP,t} - X_t \hat{\delta}_{CP,t})' \right] (\log \tilde{B}_{CP,t} - X_t \hat{\delta}_{CP,t}) + (\delta_{CP,t} - \mu_{\delta_{CP}})' X'_t X_t \left( A_{CP} + X'_t X_t \right)^{-1} A_{CP} (\delta_{CP,t} - \mu_{\delta_{CP}}) \right)
\]

\[
\sigma_{1}^{-2(q)} \left| \begin{array}{c}
\log \tilde{Y}_t(1), X_t, \mu_{\delta_1}, A_1, \alpha_1, \beta_1
\end{array} \right| \sim \Gamma \left( \alpha_1 + \frac{N_t}{2}, \beta_1 + \frac{1}{2} \left[ (\log \tilde{Y}_t(1) - X_t \hat{\delta}_{1,t})' \right] (\log \tilde{Y}_t(1) - X_t \hat{\delta}_{1,t}) + (\hat{\delta}_{1,t} - \mu_{\delta_1})' X'_t X_t \left( A_1 + X'_t X_t \right)^{-1} A_1 (\hat{\delta}_{1,t} - \mu_{\delta_1}) \right)
\]

\[
\sigma_{0}^{-2(q)} \left| \begin{array}{c}
\log \tilde{Y}_t(0), X_t, \mu_{\delta_0}, A_0, \alpha_0, \beta_0
\end{array} \right| \sim \Gamma \left( \alpha_0 + \frac{N_t}{2}, \beta_0 + \frac{1}{2} \left[ (\log \tilde{Y}_t(0) - X_t \hat{\delta}_{0,t})' \right] (\log \tilde{Y}_t(0) - X_t \hat{\delta}_{0,t}) + (\hat{\delta}_{0,t} - \mu_{\delta_0})' X'_t X_t \left( A_0 + X'_t X_t \right)^{-1} A_0 (\hat{\delta}_{0,t} - \mu_{\delta_0}) \right),
\]

where

\[
\hat{\delta}_{CP,t} = (X'_t X_t)^{-1} X'_t \log \tilde{B}_{CP,t}
\]

\[
\hat{\delta}_{1,t} = (X'_t X_t)^{-1} X'_t \log \tilde{Y}_t(1)
\]

\[
\hat{\delta}_{0,t} = (X'_t X_t)^{-1} X'_t \log \tilde{Y}_t(0),
\]
and

\[
\delta_{\text{CP}}^{(q)} \bigg| \sigma_{\text{CP}}^{2(q)} \log B_{\text{CP},t}, X_t, \mu_{\delta_{\text{CP}}}, A_{\text{CP}} \\
\sim N \left( (A_{\text{CP}} + X_t^t X_t)^{-1} \left( X_t^t \log B_{\text{CP},t} + A_{\text{CP}} \mu_{\delta_{\text{CP}}}, \right), \sigma_{\text{CP}}^{2} \left( A_{\text{CP}} + X_t^t X_t \right)^{-1} \right)
\]

\[
\delta_{1}^{(q)} \bigg| \sigma_{1}^{2(q)} \log \tilde{Y}_{t}(1), X_t, \mu_{\delta_{1}}, A_{1} \\
\sim N \left( (A_{1} + X_t^t X_t)^{-1} \left( X_t^t \log \tilde{Y}_{t}(1) + A_{1} \mu_{\delta_{1}}, \right), \sigma_{1}^{2} \left( A_{1} + X_t^t X_t \right)^{-1} \right)
\]

\[
\delta_{0}^{(q)} \bigg| \sigma_{0}^{2(q)} \log \tilde{Y}_{t}(0), X_t, \mu_{\delta_{0}}, A_{0} \\
\sim N \left( (A_{0} + X_t^t X_t)^{-1} \left( X_t^t \log \tilde{Y}_{t}(0) + A_{0} \mu_{\delta_{0}}, \right), \sigma_{0}^{2} \left( A_{0} + X_t^t X_t \right)^{-1} \right)
\]

(19)

5.4.4 Summary

The entire Gibbs sampling procedure is summarized below. If one wishes to allow for \( \rho \neq 0 \) the procedure has to be adjusted. We present this more general algorithm in the Appendix.

**Algorithm 2: Gibbs sampling**

1. Set \( \theta^{(0)} \) and \( \theta_{\text{prior}} \).
   
   for \( q = 1, \ldots, Q \) do
   
   2. Draw \( \{ \log Y_{i}^{\text{miss},(q)}(1), \log Y_{i}^{\text{miss},(q)}(0), \log B_{\text{CP},i}^{\text{miss},(q)} \} \) for all \( i = 1 \)

   3. Construct \( \{ \log \tilde{Y}_{i}^{(q)}(1), \log \tilde{Y}_{i}^{(q)}(0), \log \tilde{B}_{\text{CP},i}^{(q)} \} \) for all \( i = 1 \)

   4. Draw \( \theta^{(q)} \) according to (15)–(19).

end

5.5 Estimating optimality probability of each arm

For each draw \( \theta^{(q)} \), context \( x \) and arm \( b_{r_x} \), we can compute \( \pi^{(q)}(b_{r_x} | x) \) via equation (6). The probabilities can then be estimated by averaging over the \( Q \) draws:

\[
\hat{\psi}_{t}(b_{r_x} | x) = \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{1} \left\{ \pi^{(q)}(b_{r_x} | x, \theta^{(q)}) > \pi^{(q)}(b_{r'_x} | x, \theta^{(q)}) \right\} \text{ for all } r'_x \neq r_x.
\]

(20)
5.6 Stopping the TS

The simplest stopping rule is to specify the total number of rounds the algorithm has to run through, in which case \( c_t = t \) and \( T \) is some exogenous threshold. However, the data collected through the algorithm can help inform the decision of when to stop the experiment. We propose a stopping rule that follows this approach. We first motivate it in a non-contextual setting to provide intuition, and then generalize it to the more complex contextual case we have presented above. We then discuss the implications of using a data-based stopping rule for Bayesian inference.

5.6.1 Non-contextual case

Consider first a non-contextual MAB problem, that is, \( x \) takes one given value with probability one. Thus, we omit \( x \) for the remaining of this section to ease notation. The algorithm aims to identify the best arm while minimizing the costs of experimentation. Hence, we propose a stopping rule based on the confidence with which the optimal arm was found. More precisely, we set \( T = 0.95 \) and

\[
ct = \max_r \hat{\psi}_t(b_r),
\]

which can be interpreted as a decision to stop when the posterior distribution of \( \theta \) given \( I_t \) leads us to believe that the arm with current highest probability of being the optimal arm is the true best arm with at least 95% probability.

This stopping rule also has a well-defined interpretation in terms of Bayes factors, which are often used for Bayesian hypothesis testing. Let \( \zeta_t(b_r) \) be the posterior odds ratio of arm \( r \) being the optimal arm by the end of round \( t \). Then,

\[
\zeta_t(b_r) = \frac{\Pr_t(b_r \text{ is the optimal bid})}{\Pr_t(b_r \text{ is not the optimal bid})} = \frac{\Pr_t(b_r \text{ is the optimal bid})}{1 - \Pr_t(b_r \text{ is the optimal bid})} = \frac{\psi_t(b_r)}{1 - \psi_t(b_r)}.
\]

Thus, \( ct \) can alternatively be constructed as \( \max_r \zeta_t(b_r) \), with corresponding threshold
$T = 19$, so that stopping is based off a threshold on the implied Bayes factor.$^6$

5.6.2 Contextual case

Having provided the motivation behind our proposed stopping criterion, we now proceed to adapt it to the contextual MAB problem. This is a more complex problem because now there is not a single best arm, but $p$ best arms. Thus, a natural but conservative approach would be to require 95% posterior probability over a list of $p$ arms as being the optimal ones. In this case, the threshold rule can be expressed by:

$$c_t = \min_{x \in X} \max_{r_s} \hat{\psi}_t(b_{r_s} | x),$$

while maintaining the requirement that $c_t > T = 0.95$. Consequently, upon stoppage there would be at least 95% posterior probability on a given $b^*(x)$ for each $x$.

In some scenarios it is arguably the case that the advertiser’s inference objective is to estimate and perform inference on the unconditional $ATE$. Under these circumstances, the stopping rule above is likely to be too stringent since the goal is not to learn every $ATE(x)$ with high probability, but just $ATE$. Hence, we now present a slightly less demanding approach.

Recall that for each context $x$ we consider $R_x$ different arms. Consequently, the unconditional $ATE$ can take at most $R = \prod_{\ell=1}^p R_{x_\ell}$ values as $ATE = \sum_{\ell=1}^p F_x(x_\ell) \times ATE(x_\ell)$. Consider a grid with the $S$ unique values among these $R$, which we denote $a_s$ for $s = 1, ..., S$. Our criterion is to stop when at least 95% of the Q draws from the posterior imply that the $ATE$ is equal to one of the $a_s$ values in the grid.

To make this criterion precise, notice that, at the end of round $t$, for each $\theta^{(q)}$ and context $x$, there is an arm that maximizes $\pi \left( \cdot | x, \theta^{(q)} \right)$. In turn, the bid associated with such an arm is equal to $ATE(x)$ by Proposition 1.$^7$ Denote this value by $ATE^{(q)}_t(x)$. It is straightforward to then compute $ATE^{(q)}_{t} = \sum_{\ell=1}^p F_x(x_\ell) \times ATE^{(q)}_{t}(x_\ell)$. Our criterion is

---

$^6$It is important to emphasize that following this stopping rule is not equivalent to conducting a sequential Bayesian hypothesis test. Such procedure would require us to establish a null hypothesis that one specific arm was the best, and base the decision to stop solely on this arm’s Bayes factor or posterior odds ratio. Instead, we remain agnostic as to which arm is the best, and base our decision to stop on which arm has strongest evidence on its support.

$^7$Hence, we implicitly assume that all the CATEs are non-negative; otherwise, the resulting estimate of the $ATE$ should be seen as an upper bound of the true $ATE$. 

20
then given by:

\[ c_t = \max_{s \in \{1, \ldots, S\}} \frac{1}{Q} \sum_{q=1}^{Q} 1 \{ ATE_t^{(q)} = a_s \}. \]  

(24)

Notice that while our decision rule depends on the confidence with which we believe to have found the true unconditional ATE as implied by the posterior distribution of \( \theta \) given \( I_t \), traffic is still allocated to each arm according to (20). Hence, the decision to stop the experiment is aligned with the advertiser’s inference goal, while the way it performs randomization is aligned with her economic goal. Finally, note that this stopping criterion presupposes that the distribution from which contexts are drawn, \( F_x(\cdot) \), is known to the researcher. When this is not the case, one could replace it with empirical frequencies estimated using data collected via the algorithm, with the slight added difficulty that the grid of unique \( S \) values \( ATE_t \) can take now will potentially change round by round. While we expect the stopping rule given in (24) to shorten the duration of the experiment when compared to the one given in (23), we found in simulations that the difference between these two rules is minimal.

### 5.6.3 Implications for Bayesian inference

It is important to mention that the question of how to stop an experiment while conducting Bayesian inference is still an open issue. Even though data-based stopping rules are known to interfere with frequentist inference, which motivated the development of new methods to explicitly account for this interference both for non-adaptive (Johari et al., 2016) and adaptive (Yang et al., 2017; Jamieson and Jain, 2018; Ju et al., 2019) data collection procedures, Bayesian inference has historically been viewed as immune to optional stopping rules (Lindley, 1957; Edwards et al., 1963; Savage, 1972; Good, 1991). Nevertheless, a recent debate has emerged concerning what are the effects of optional stopping on frequentist properties and interpretation of Bayes estimators (Yu et al., 2014; Sanborn and Hills, 2014; Rouder, 2014; Dienes, 2016; Deng et al., 2016; Schönbrodt et al., 2017; Hendriksen et al., 2018; de Heide and Grünwald, 2018; Wagenmakers et al., 2019; Rouder, 2019). We do not attempt to resolve this debate in this paper. Instead, we present simulations that show its practical impact is minimal in our setting.
5.7 Practical considerations

5.7.1 Expected regret minimization versus best arm identification

We set up the problem in an expected regret minimization framework based on the viewpoint that the advertiser seeks to maximize his payoffs from bidding during the experiment. We could alternatively cast the problem as one of pure best arm identification (Bubeck et al., 2009).\textsuperscript{8} To leverage Proposition 1, all we need is a bandit framework to recover the arm with largest expected reward, so the core idea behind the approach ports to that situation as well.

5.7.2 Parametric assumptions

More flexible parametric specifications can be used instead of (5) and (7) if the researcher is willing to employ more complex MCMC methods, which we expect to be slower than the Gibbs sampling algorithm presented above. This is because conditional conjugacy is likely to fail under alternative parametrizations. However, any algorithm with convergence guarantees (e.g., UCB, $\epsilon$-greedy) could in theory be used instead of the TS algorithm if the practitioner is not comfortable with making specific distributional assumptions, which are not required for these other methods.

5.7.3 Obtaining draws from posterior distribution

The method we presented requires the researcher to employ Gibbs sampling on an increasing data set every round, which becomes slower as the number of observations increases. We note that Sequential Monte Carlo (SMC) methods could instead be used to speed up the sampling.\textsuperscript{9}

---

\textsuperscript{8}For an example of a study that adopts this approach to identify an optimal policy see Kasy and Sautmann (2019).

\textsuperscript{9}For an application of SMC methods to MAB problems, see Cherkassky and Bornn (2013).
5.7.4 Using additional data on competing bids

We also note that it is straightforward to adapt the procedure depending on which auction data are made available. We have assumed that the advertiser only observes $B_{CP}$ when she effectively has to pay this amount; otherwise, all she knows is that it is bounded below by $b$. If the transaction price from the auction is made public, the advertiser obtains a more precise lower bound whenever it is larger than $b$. This yields a new definition of $\bar{B}_{CP}$ and does not require any modification of the procedure. However, if $B_{CP}$ itself is made public the algorithm simplifies since the censoring problem vanishes. Hence, the practitioner can update the posterior distribution over the parameters $\delta_{CP}$ and $\sigma_{CP}^2$ analytically, without the need to use the Gibbs sampling procedure introduced by Chib (1992), because of the exact conjugacy implied by (5) and (7).

5.7.5 Addressing the cold start problem

While priors always play an important role in Bayesian inference, they can become even more important in the context of experimentation as a way to deal with the cold start problem. Well-informed priors might situate the algorithm at a good starting point, shortening the duration of the experiment and, consequently, decreasing its costs. On the other hand, poorly specified priors might have the opposite effect and become inferior even to diffuse, non-informative priors. Hence, we now propose an approach that uses past data to inform the choice of the parameters $\{\mu_{\delta_k}, A_k, a_k, \beta_k\}_{k \in \{1,0, CP\}}$. Throughout this section, we assume that the researcher has access to a data set $I_n = \{b_i, X_i, D_i, Y_i, \bar{B}_{CP}, i \}_{i=1}^n$.

Our approach is based on equating the means and variances of the prior distributions to the approximate means and variances of estimators of $\delta_1$, $\delta_0$, $\delta_{CP}$ and $\sigma^2 \equiv [\sigma_1^2, \sigma_0^2, \sigma_{CP}^2]'$. Since we employ maximum likelihood estimators (MLEs), which in this

\footnote{Feit and Berman (2019) also estimate priors from historical data.}
case are \(\sqrt{n}\)-consistent and asymptotically normal, our approach sets, for \(k \in \{1, 0, CP\}\),

\[
\frac{\alpha_k}{\beta_k} = \hat{\sigma}_k^{-2} \\
\frac{\alpha_k}{\beta_k^2} = \frac{1}{n}A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\sigma}_k^{-2} - \sigma_k^{-2} \right) \right] = \frac{\hat{\sigma}_k^8}{n}A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\sigma}_k^2 - \sigma_k^2 \right) \right] \\
\mu_k = \hat{\delta}_k \\
A_k = n\hat{\sigma}_k^2 \left\{ A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\delta}_k - \delta_k \right) \right] \right\}^{-1}.
\] (25)

Notice that for the parameters of the Gamma distributions this is equivalent to:

\[
\alpha_k = n\hat{\sigma}_k^{-4} \left\{ A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\delta}_k^{-2} - \sigma_k^{-2} \right) \right] \right\}^{-1} = n\hat{\sigma}_k^4 \left\{ A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\delta}_k^2 - \sigma_k^2 \right) \right] \right\}^{-1} \\
\beta_k = n\hat{\sigma}_k^{-2} \left\{ A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\delta}_k^{-2} - \sigma_k^{-2} \right) \right] \right\}^{-1} = n\hat{\sigma}_k^6 \left\{ A\hat{\text{var}} \left[ \sqrt{n} \left( \hat{\delta}_k^2 - \sigma_k^2 \right) \right] \right\}^{-1}.
\] (26)

For the purposes of estimation we assume that \(D_i \perp \perp Y_i(1), Y_i(0)|X_i\). Because of Assumption 3, the only potential source of dependence between \(D_i\) and the potential outcomes is \(b_i\). This is not a concern when the algorithm is implemented since the bids \(b_i\) are under the control of the researcher. If the historical data come from an experiment in which bids were randomized, this condition is also satisfied.\(^{11}\) However, we note that the algorithm will consistently recover the true CATEs even if this assumption is violated and the resulting priors are poorly specified. We provide details about the estimators and their computation in the Appendix.

6 Simulations

To illustrate our method we perform a series of simulations. We consider a setting with two equiprobable contexts and set \(\delta_1 \approx [0.81, 1.25]', \delta_0 = [0.2, 0.45]', \delta_{CP} = [0.25, 0.4]', \sigma_1^2 = 0.36, \sigma_0^2 = 0.64, \sigma_{CP}^2 = 0.25\) and \(\rho = 0\), so that \(b^*(x_1) = ATE(x_1) = 1, b^*(x_2) = ATE(x_2) = 2\) and \(ATE = 1.5\).

A simple way to estimate the ATEs would be to perform a non-adaptive experiment in which users are randomly exposed to the ad, which we refer to as an “A/B test”. This is not possible in RTB since the advertiser does not fully control the ad delivery mechanism.

\(^{11}\)Hence, Assumption 3 can also be seen as a version of unconfoundedness (also known as conditional independence and selection on observables) from the econometrics and statistics literatures.
However, by randomly picking arbitrarily high and low bids the advertiser can virtually guarantee to win or lose the auction, thereby obtaining random ad-exposure. We equate these high and low bids to 20 and 0.01 for both contexts and consider three randomization scenarios: $Pr(D = 1) = Pr(b = 20) \in \{0.25, 0.5, 0.75\}$. Figure 2a plots the expected regret per round of each of these randomization policies over a sequence of 100 rounds of bidding, which is the expected profit from equiprobably bidding 20 or 0.01. Since these strategies are non-adaptive, the expected regret per round is constant.

We then consider a MAB, with bids $\{0.5, 1.0, 1.5\}$ for context $x_1$ and $\{1, 2, 3\}$ for context $x_2$. We run 1,000 simulations using our BTS algorithm, each of which for 100 rounds with 50 new auctions per round, so that $n_t = 50, 1 \leq t \leq 100$. We set the following parameters for the priors: $\{\mu_k = A_k = \alpha_k = \beta_k = 0\}_{k \in \{1,0,CP\}}$. The Gibbs sampling always started at the values $\{\delta_k = 0\}_{k \in \{1,0,CP\}}$ and $\{\sigma^2_k = 1\}_{k \in \{1,0,CP\}}'$ and the initial optimality probabilities were set to $\psi_0(b|x_1) = 1/3$ for $b \in \{0.5, 1.0, 1.5\}$ and $\psi_0(b|x_2) = 1/3$ for $b \in \{1, 2, 3\}$. On each round, we drew 1,000 vectors of parameters from the posterior distribution, dropped the first 250 and used the remaining 750 to estimate $\psi_t(\cdot|\cdot)$.

To compare this approach to the simpler non-adaptive experimentation policies, we also display in Figure 2a boxplots with the interquartiles of expected regret from the BTS algorithm for each round across the 1,000 simulations. That is, at the end of round $t$ we compute the expected profit from bidding according to the probabilities $\hat{\psi}_t(\cdot|\cdot)$, which we denote by $\bar{\pi}_t(\cdot)$, and subtract it from $\bar{\pi}(b^*) = \sum_{\ell=1}^2 F(x_\ell) \times \bar{\pi}(b^*(x_\ell)|x_\ell)$. The results show that the BTS algorithm dominates the simpler randomization policies in terms of expected regret per round.

The fact that the expected regret per round from the BTS algorithm tends to zero suggests that this method succeeds in identifying the true best arm. To better illustrate this, Figure 2b displays boxplots with the interquartiles of $\hat{Pr}(ATE_t = ATE)$ across the 1,000 simulations. Results show that as the algorithm progresses and data are accumulated, the interquartile plots uniformly converge to 1, demonstrating that our method succeeds in identifying the true $ATE$s in practice.
Figure 2: Comparing BTS to A/B tests: interquartiles over 100 rounds across 1,000 simulations
The comparison between the BTS algorithm and the extreme A/B tests described above might stack the deck in favor of the MAB approach. To address this concern we recreate Figure 2a but this time comparing the MAB approach to an A/B test implemented on the same grid of bids considered by the BTS algorithm, that is, one in which each of three bids from the grids is chosen with equal probability. We display the results in Figure 2c. Despite being less drastic, they still show that the BTS algorithm dominates the non-adaptive randomization policy in terms of expected regret per round.

The dominance of the BTS algorithm over the A/B test on the grid of bids is more intense if we consider the evolution of cumulative expected regret, which is given by \( t \times \bar{\pi}(b^*) - \sum_{t=1}^{T} \bar{\pi}_t(\cdot) \), instead of expected regret per round. This is a relevant quantity in situations where the advertiser has a fixed amount to spend on an experiment and wishes to obtain as much data as possible with this amount. We thus replicate Figure 2c but displaying the cumulative expected regret over a sequence of 100 rounds across the 1,000 simulations. Figure 2d shows that the MAB approach allows the advertiser to collect more data than a non-adaptive randomization policy, which is arguably desirable for the purposes of accomplishing her dual objective.

The aforementioned results ignore the stopping rule we introduced in section 5.6. Figure 2b showed that \( \hat{\text{Pr}}(ATE_t = ATE) \) always converged to one as the algorithm progressed. However, our proposed stopping rule stipulates that the algorithm should stop once at least 95% of the posterior optimality probability concentrated on any of the possible values \( ATE_t \) can take, which would then be identified as the true \( ATE \). Thus, the relevant question is twofold: first, would the algorithm indeed have stopped before the hundredth round had the stopping rule been applied, and second, would it have correctly identified the true \( ATE \) when it stopped.

Over the 1,000 simulations, the algorithm would have stopped before the hundredth round in all but 54 times and would have identified the true \( ATE \) upon stoppage in 97.57% of such cases. Figure 3 presents a histogram of the stopping times associated with the simulations from Figure 2 that would have stopped before the hundredth round had the stopping rule from section 5.6 been used. Stopping times associated with correct identification of the true \( ATE \) are displayed in blue, while the ones associated with mistakes are shown in red. Results show that the algorithm is much more likely to make a mistake when it stops at very early stages: more than 50% of mistakes happened when the algorithm ran for at most 21 rounds, while the minimum amount it needed to correctly identify the true \( ATE \) at stoppage was 22. Hence, in practice we suggest incorporating into the stopping
rule a restriction that the algorithm runs for a minimum number of rounds.

7 Concluding remarks

An online algorithm for obtaining causal inference on RTB advertising is presented. The algorithm leverages the theory of optimal bidding under second-price sealed-bid auctions to align the twin goals of obtaining economic payoff maximization and inference on the expected effect of advertising. The algorithm is implemented as a modified TS that is adaptively updated via MCMC. The second-price sealed-bid auction environment is the most popular auction format for RTB ads (see, e.g., Choi et al., 2017). Extensions to more complex auction environments (e.g., first-price sealed-bid auctions) can make the inference problem more challenging. As mentioned before, allowing for contexts will make Assumption 3 more viable by leveraging a conditional independence assumption. Logged historical data can be used to develop data-driven priors and solve the cold-start problem. These extensions are being pursued in our future work as part of the implementation of the algorithm on the advertising platform of JD.com.
References

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, 24(1–2):3–61.

Athey, S. and Wager, S. (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*.

Austin, D., Seljan, S., Moreno, J., and Tzeng, S. (2016). Reserve price optimization at scale. In *DSAA 2016*, pages 528–536.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In *NIPS 2015*, pages 1342–1350.

Bastani, H. and Bayati, M. (2015). Online decision-making with high-dimensional covariates. *Working paper, Stanford University*.

Bergemann, D. and Välimäki, J. (2008). Bandit problems. In Durlauf, S. N. and Blume, L. E., editors, *The New Palgrave Dictionary of Economics: Volume 1 – 8*, pages 336–340. Palgrave Macmillan UK, London.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *ALT 2009*, pages 23–37.

Cai, H., Ren, K., Zhang, W., Malialis, K., Wang, J., Yu, Y., and Guo, D. (2017). Real-time bidding by reinforcement learning in display advertising. In *WSDM 2017*, pages 661–670.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2014). Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564.

Chawla, S., Hartline, J., and Nekipelov, D. (2016). A/B testing of auctions. *arXiv preprint arXiv:1606.00908*.

Cherkassky, M. and Bornn, L. (2013). Sequential Monte Carlo bandits. *arXiv preprint arXiv:1310.1404*.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51(1–2):79–99.
Choi, H., Mela, C., Balseiro, S., and Leary, A. (2017). Online display advertising markets: A literature review and future directions. Working paper, Duke University.

de Heide, R. and Grünwald, P. D. (2018). Why optional stopping is a problem for Bayesians. arXiv preprint arXiv:1708.08278.

Deng, A., Lu, J., and Chen, S. (2016). Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In DSAA 2016, pages 243–252.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. (2018). Accurate inference for adaptive linear models. In ICML 2018, pages 1194–1203.

Diemert, E., Héliou, A., and Renaudin, C. (2018). Off-policy learning for causal advertising. NIPS 2018, Workshop on Causal Learning.

Dienes, Z. (2016). How Bayes factors change scientific practice. Journal of Mathematical Psychology, 72:78–89.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2018). Estimation considerations in contextual bandits. arXiv preprint arXiv:1711.07077.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In ICML 2011, pages 1097–1104.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70(3):193–242.

Feit, E. M. and Berman, R. (2019). Test & roll: Profit-maximizing A/B tests. Marketing Science, Forthcoming.

Forney, A., Pearl, J., and Bareinboim, E. (2017). Counterfactual data-fusion for online reinforcement learners. In ICML 2017, pages 1156–1164.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. Stochastic Systems, 3(1):230–261.

Good, I. J. (1991). A comment concerning optional stopping. Journal of Statistical Computation and Simulation, 39(3):191–192.

Hendriksen, A., de Heide, R., and Grünwald, P. D. (2018). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. arXiv preprint arXiv:1807.09077.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327.

Jamieson, K. G. and Jain, L. (2018). A bandit approach to sequential experimental design with false discovery control. In *NIPS 2018*, pages 3664–3674.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *ICML 2016*, pages 652–661.

Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. (2018). Real-time bidding with multi-agent reinforcement learning in display advertising. In *CIKM 2018*, pages 2193–2201.

Johari, R., Pekelis, L., and Walsh, D. J. (2016). Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*.

Ju, N., Hu, D., Henderson, A., and Hong, L. (2019). A sequential test for selecting the better variant: Online A/B testing, adaptive allocation, and continuous monitoring. In *WSDM 2019*, pages 492–500.

Kallus, N. (2018a). Balanced policy evaluation and learning. In *NIPS 2018*, pages 8895–8906.

Kallus, N. (2018b). Instrument-armed bandits. In *ALT 2018*, pages 529–546.

Kasy, M. and Sautmann, A. (2019). Adaptive treatment assignment in experiments for policy choice. *Working paper, Harvard University*.

Koop, G. and Poirier, D. J. (1997). Learning about the across-regime correlation in switching regression models. *Journal of Econometrics*, 78(2):217–227.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian Econometric Methods*. Cambridge University Press.

Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. In *NIPS 2016*, pages 1181–1189.

Lewis, R. and Wong, J. (2018). Incrementality bidding & attribution. *Working paper*. 
Li, L., Chen, S., Kleban, J., and Gupta, A. (2015). Counterfactual estimation and optimization of click metrics in search engines: A case study. In WWW 2015, pages 929–934.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):187–192.

Misra, K., Schwartz, E. M., and Abernethy, J. (2019). Dynamic online pricing with incomplete information using multi-armed bandit experiments. *Marketing Science*, 38(2):226–252.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2018). Why adaptively collected data have negative bias and how to correct for it. In AISTATS 2018, pages 1261–1269.

Olsen, R. J. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica*, 46(5):1211–1215.

Ostrovsky, M. and Schwarz, M. (2016). Reserve prices in internet advertising auctions: A field experiment. Working paper, Stanford University.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pouget-Abadie, J., Mirrokni, V., Parkes, D. C., and Airoldi, E. M. (2018). Optimizing cluster-based randomized experiments under monotonicity. In SIGKDD 2018, pages 2090–2099. ACM.

Rhuggenaath, J., Akcay, A., Zhang, Y., and Kaymak, U. (2019). Optimizing reserve prices for publishers in online ad auctions. In CIFEr 2019.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2):301–308.

Rouder, J. N. (2019). On the interpretation of Bayes factors: A reply to de Heide and Grünwald. *PsyArXiv Preprints*.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sanborn, A. N. and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, 21(2):283–300.

Savage, L. J. (1972). *The Foundations of Statistics*. Courier Corporation.
Sawant, N., Namballa, C. B., Sadagopan, N., and Nassif, H. (2018). Contextual multi-armed bandits for causal marketing. *ICML 2018, Workshop on Causal ML*.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2):322–339.

Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry, 26*(6):639–658.

Strehl, A., Langford, J., Li, L., and Kakade, S. M. (2010). Learning from logged implicit exploration data. In *NIPS 2010*, pages 2217–2225.

Swaminathan, A. and Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML 2015*, pages 814–823.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML 2016*, pages 2139–2148.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25*(3/4):285–294.

Varian, H. R. (2009). Online ad auctions. *American Economic Review P & P, 99*(2):430–34.

Vijverberg, W. P. M. (1993). Measuring the unidentified parameter of the extended Roy model of selectivity. *Journal of Econometrics, 57*(1–3):69–89.

Villar, S. S., Bowden, J., and Wason, J. M. S. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science, 30*(2):199–215.

Wagenmakers, E.-J., Gronau, Q. F., and Vandekerckhove, J. (2019). Five Bayesian intuitions for the stopping rule principle. *PsyArXiv Preprints*.

Wang, Y., Liu, J., Liu, Y., Hao, J., He, Y., Hu, J., Yan, W. P., and Li, M. (2017a). LADDER: A human-level bidding agent for large-scale real-time online auctions. *arXiv preprint arXiv:1708.05565*.

Wang, Y.-X., Agarwal, A., and Dudík, M. (2017b). Optimal and adaptive off-policy evaluation in contextual bandits. In *ICML 2017*, pages 3589–3597.
Wu, D., Chen, C., Yang, X., Chen, X., Tan, Q., Xu, J., and Gai, K. (2018). A multi-agent reinforcement learning method for impression allocation in online display advertising. *arXiv preprint arXiv:1809.03152*.

Xu, M., Qin, T., and Liu, T.-Y. (2013). Estimation bias in multi-armed bandit algorithms for search advertising. In *NIPS 2013*, pages 2400–2408.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. (2017). A framework for Multi-A(rmed)/B(andit) testing with online FDR control. In *NIPS 2017*, pages 5957–5966.

Yu, E. C., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, 21(2):268–282.

Yuan, S., Wang, J., and Zhao, X. (2013). Real-time bidding for online advertising: Measurement and analysis. In *KDD 2013*, pages 1–8.
Appendix

A Maximum likelihood estimators used on historical data

This section describes in more detail the maximum likelihood estimators (MLEs) we use to choose the parameters of the prior distributions. The assumptions we make on the historical data are the same as the ones discussed in section 5.3.

A.1 Potential outcomes

We begin by describing how we use historical data to pick the parameters for the prior distributions associated with the potential outcomes. Because we are maintaining the assumption that $D_i \perp Y_i(1), Y_i(0) \mid X$, we can simply use OLS on historical data to pick the parameters since it is equivalent to the MLE. In particular, define $X_{i1} \equiv D_i X_i$. It follows that:

$$
\hat{\delta}_1 = \left(\frac{1}{n} \sum_{i=1}^{n} X_{i1} X_{i1}'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i1} \log Y_i\right),
$$

$$
\hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log Y_i - X_{i1}' \hat{\delta}_1\right)^2
$$

and

$$
\sqrt{n} \begin{bmatrix} \hat{\delta}_1 - \delta_1 \\ \hat{\sigma}_1^2 - \sigma_1^2 \end{bmatrix} \xrightarrow{d} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 (\mathbb{E} [X_{i1} X_{i1}'])^{-1} & 0 \\ 0 & 2\sigma_1^2 \end{bmatrix} \right).
$$

Hence,

$$
\text{Avar} \left[\sqrt{n} (\hat{\delta}_1 - \delta_1)\right] = \hat{\sigma}_1^2 \left(\frac{1}{n} \sum_{i=1}^{n} X_{i1} X_{i1}'\right)^{-1}
$$

and

$$
\text{Avar} \left[\sqrt{n} (\hat{\sigma}_1^2 - \sigma_1^2)\right] = 2\hat{\sigma}_1^4.
$$
The estimators $\hat{\delta}_0$ and $\hat{\sigma}_0^2$ are analogous to the ones above, with $X_{i0} \equiv (1 - D_i)X_i$ replacing $X_{i1}$, so we omit them for brevity.

### A.2 Highest competing bid

Even though we maintain the assumption of treatment exogeneity, we still have to account for censoring of the highest competing bid. Given the normality assumption, the censoring characterizes a standard Tobit model. To make its MLE more computationally manageable, we first reparametrize the model so that the log-likelihood function becomes globally concave as first shown by Olsen (1978). Let $\nu_{CP} \equiv \sigma_{CP}^{-1} \delta_{CP}$ and $\tau_{CP} \equiv \sigma_{CP}^{-1}$. The log-likelihood of the data is then given by:

$$
\log L (I_n | \nu_{CP}, \tau_{CP}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ D_i \log \left[ \tau_{CP} \phi \left( \tau_{CP} \log b_{CP,i} - X_i' \nu_{CP} \right) \right] \\
+ (1 - D_i) \log \left[ \Phi \left( X_i' \nu_{CP} - \tau_{CP} \log b_i \right) \right] \right\},
$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively.

We use the Newton-Raphson algorithm to compute the estimator. This requires us to compute the first and second derivatives of the log-likelihood function. We have that:
\[
\frac{\partial \log L}{\partial \theta_{CP}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ D_i \left( \theta_{CP} \log b_{CP,i} - X'_i \theta_{CP} \right) + (1 - D_i) \frac{\phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)}{\Phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)} \right\} X_i
\]

\[
\frac{\partial^2 \log L}{\partial \theta_{CP} \partial \theta_{CP}'} = \frac{1}{n} \sum_{i=1}^{n} \left\{ D_i \left( \frac{1}{\theta_{CP}} - \log b_{CP,i} \left( \theta_{CP} \log b_{CP,i} - X'_i \theta_{CP} \right) \right) \right\} X_i X'_i
\]

\[
\frac{\partial^2 \log L}{\partial \theta_{CP} \partial \theta_{CP}'} = \frac{1}{n} \sum_{i=1}^{n} \left\{ D_i \log b_{CP,i} + (1 - D_i) \frac{\phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)}{\Phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)} \right\} \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)
\]

\[
\frac{\partial^2 \log L}{\partial \theta_{CP}^2} = -\frac{1}{n} \sum_{i=1}^{n} \left\{ D_i \left( \frac{1}{\theta_{CP}} + (\log b_{CP,i})^2 \right) - (1 - D_i) \frac{\phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)}{\Phi \left( X'_i \theta_{CP} - \theta_{CP} \log b_i \right)} \right\} \left( \log b_i \right)^2
\]

Letting

\[
H \left( \theta_{CP}, \theta_{CP} \right) = \begin{bmatrix}
\frac{\partial^2 \log L(\theta_{CP}, \theta_{CP})}{\partial \theta_{CP} \partial \theta_{CP}'} & \frac{\partial^2 \log L(\theta_{CP}, \theta_{CP})}{\partial \theta_{CP} \partial \theta_{CP}'}
\end{bmatrix},
\]

it then follows that

\[
\sqrt{n} \left[ \hat{\theta}_{CP} - \theta_{CP} \right] \xrightarrow{d} N \left( 0, \text{plim}_{n \to \infty} \left[ H \left( \theta_{CP}, \theta_{CP} \right)^{-1} \right] \right).
\]

To convert the parameters back to the original ones, we make use of the delta method.
We have that:
\[
g(\mathcal{N}_{CP}, \mathcal{D}_{CP}) = \left[ \begin{array}{c} -1 \mathcal{N}_{CP} \\ -2 \mathcal{D}_{CP} \end{array} \right],
\]
which implies that
\[
\nabla g(\mathcal{N}_{CP}, \mathcal{D}_{CP}) = \left[ \begin{array}{cc} \frac{\partial g(\mathcal{N}_{CP}, \mathcal{D}_{CP})}{\partial \mathcal{N}_{CP}} & \frac{\partial g(\mathcal{N}_{CP}, \mathcal{D}_{CP})}{\partial \mathcal{D}_{CP}} \\ \partial \mathcal{N}_{CP} & \partial \mathcal{D}_{CP} \end{array} \right] = \left[ \begin{array}{cc} -1 \mathcal{N}_{CP} I_p & -2 \mathcal{D}_{CP} \mathcal{N}_{CP} \\ 0 & -3 \mathcal{D}_{CP} \end{array} \right],
\]
where \( I_p \) is the identity matrix with dimension \( p \). Thus, by the delta method:
\[
\sqrt{n} \left[ \delta_{CP} - \delta_{CP} \right] \overset{d}{\rightarrow} N \left( 0, -\lim_{n \to \infty} \left[ \nabla g(\hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP}) H \left( \hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP} \right)^{-1} \nabla g(\hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP})' \right] \right).
\]
Finally, \( \hat{A}\text{var} \left[ \sqrt{n} \left( \delta_{CP} - \delta_{CP} \right) \right] \) and \( \hat{A}\text{var} \left[ \sqrt{n} \left( \delta_{CP}^2 - \sigma_{CP}^2 \right) \right] \) are obtained by picking the block diagonal elements of the matrix
\[
-\nabla g(\hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP}) H \left( \hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP} \right)^{-1} \nabla g(\hat{\mathcal{N}}_{CP}, \hat{\mathcal{D}}_{CP})'.
\]

B Gibbs sampling when potential outcomes are correlated

We now present a more general Gibbs sampling that accommodates the possibility that \( \rho \neq 0 \). When \( \rho \neq 0 \), the missing values \( \log Y_{i}^{\text{miss}} \) depend on the observed values \( \log Y_{i} \) even conditional on \( D_{i} \), which requires us to change the priors and the procedure accordingly. To do so, we combine the Bayesian estimator for the standard Tobit model introduced by Chib (1992) and the approach to estimate the parameters in a seemingly unrelated regressions (SUR) model where all equations have the same set of regressors with data augmentation in a single Gibbs sampling algorithm.\(^{12}\) We now present these adaptations in detail.

\(^{12}\)See section 2.8.5 of Rossi et al. (2005) and section 14.11 of Koop et al. (2007) for more details.
B.1 Prior distributions

For \( k \in \{1, 0\} \) we replace (7) with

\[
\Sigma^{-1} \equiv \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_0 \\ \rho \sigma_1 \sigma_0 & \sigma_0^2 \end{bmatrix}^{-1} \sim W(\nu, \Xi^{-1})
\]

\[
\delta \equiv \text{vec}(\Delta) = \begin{bmatrix} \delta_1 \\ \delta_0 \end{bmatrix} \sim N\left(\mu_{\delta}, \Sigma \otimes A_{\delta}^{-1}\right)
\]

(B.1)

where \( W(\cdot, \cdot) \) denotes the Wishart distribution, \( \nu \) is a non-negative scalar, \( \Xi \) is a 2-by-2 matrix, \( \mu_{\delta} = \left[\mu_{\delta_1}', \mu_{\delta_0}'\right]' \) is a 2\( p \)-by-1 vector and \( A_{\delta} \) is a \( p \)-by-\( p \) matrix.\(^{13}\) We will also use the following \( p \)-by-2 matrix: \( M_{\delta} \equiv \left[\mu_{\delta_1}, \mu_{\delta_0}\right] \).

B.2 Distributions of missing values, data augmentation and completion

Instead of (9) and (10) it now follows that:

\[
\log Y_{i,m}^{\text{miss}}(1) \big| D_i = 0, \log Y_i, \log \bar{B}_{CP,i}, \log b_i, X_i, \theta \overset{d}{=} \log Y_{i,m}^{\text{miss}}(1) \big| D_i = 0, \log Y_i, X_i, \delta, \Sigma \sim N\left(X_i'\delta_1 + \frac{\rho \sigma_1}{\sigma_0} (\log Y_i - X_i'\delta_0), (1 - \rho^2) \sigma_1^2\right)
\]

(B.2)

and

\[
\log Y_{i,m}^{\text{miss}}(0) \big| D_i = 1, \log Y_i, \log \bar{B}_{CP,i}, \log b_i, X_i, \theta \overset{d}{=} \log Y_{i,m}^{\text{miss}}(0) \big| D_i = 1, \log Y_i, X_i, \delta, \Sigma \sim N\left(X_i'\delta_0 + \frac{\rho \sigma_0}{\sigma_1} (\log Y_i - X_i'\delta_1), (1 - \rho^2) \sigma_0^2\right),
\]

(B.3)

while (8) remains the same. We can redefine \( \delta_{i,m}^{\text{miss}} \) and \( \sigma_{i,2,m}^{\text{miss}} \) as

\[
\delta_{i,m}^{\text{miss}} = D_i \left(X_i'\delta_0 + \frac{\rho \sigma_0}{\sigma_1} (\log Y_i - X_i'\delta_1)\right) + (1 - D_i) \left(X_i'\delta_1 + \frac{\rho \sigma_1}{\sigma_0} (\log Y_i - X_i'\delta_0)\right)
\]

(B.4)

\(^{13}\)We maintain independent priors for the parameters associated with \( \{Y(1), Y(0)\} \) and \( B_{CP} \) because of Assumption 3. Should this assumption be relaxed, we could then express (7) including \( \delta_{CP} \) into \( \delta \) and \( \Delta \) and the same for \( \sigma_{CP}^2 \) and the correlations between \( \log B_{CP} \) and \( \log Y(1) \) and \( \log Y(0) \) into the matrix \( \Sigma \).
and

\[ \sigma_{i,\text{miss}}^2 = (1 - \rho^2) \left[ D_i \sigma_0^2 + (1 - D_i) \sigma_1^2 \right], \]  

(B.5)

respectively, and combine them into

\[ \log Y_i^{\text{miss}} \mid \log Y_i, D_i, X_i, \delta, \Sigma \sim N \left( \delta_i^{\text{miss}}, \sigma_i^{2,\text{miss}} \right). \]  

(B.6)

The completion process in (14) remains unchanged.

### B.3 Drawing from posterior distribution

We once again condition on the “completed” data, \( \tilde{I}_t \), and on the parameters of the prior distributions, which now are given by \( \theta_{\text{prior}} \equiv \{ \mu_\delta, A_\delta, \nu, \Xi, \mu_{\delta_{CP}}, A_{CP}, \alpha_{CP}, \beta_{CP} \} \). In addition to the previously defined objects, we will also use the following \( N_t \)-by-2 matrix,

\[ \log \tilde{Y}_{PO,t} \equiv \begin{bmatrix} \log \tilde{Y}_{t} \mid \quad 1 \\ \log \tilde{Y}_{t} \mid \quad 0 \end{bmatrix}, \]

and

\[ \tilde{\Delta}_t = (X'_t X_t + A_\delta)^{-1} (X'_t \log \tilde{Y}_{PO,t} + A_\delta M_\delta) \]  

(B.7)

and

\[ SSR_t = (\log \tilde{Y}_{PO,t} - X_t \tilde{\Delta}_t)' (\log \tilde{Y}_{PO,t} - X_t \tilde{\Delta}_t) + (\tilde{\Delta}_t - M_\delta)' A_\delta (\tilde{\Delta}_t - M_\delta). \]  

(B.8)

To draw new values for \( \sigma_{CP}^2 \) and \( \delta_{CP} \) we still utilize expressions (17) and (19). However, instead of using these expressions to draw new values for \( \Sigma \) and \( \delta \), we now leverage the following results:

\[ \Sigma^{-1, (q)} \mid \theta^{(q-1)}, \theta_{\text{prior}}, \tilde{I}_t \overset{d}{=} \Sigma^{-1, (q)} \mid \log \tilde{Y}_{PO,t}, X_t, \nu, \Xi, \mu_\delta, A_\delta \]  

(B.9)

and

\[ \delta^{(q)} \mid \Sigma^{(q)}, \sigma_{CP}^{2, (q)}, \delta^{(q-1)}, \delta_{CP}^{(q-1)}, \theta_{\text{prior}}, \tilde{I}_t \overset{d}{=} \delta^{(q)} \mid \Sigma^{(q)}, \log \tilde{Y}_{PO,t}, X_t, \mu_\delta, A_\delta. \]  

(B.10)

For completeness, given the parametric assumptions we made it follows that:

\[ \Sigma^{-1, (q)} \mid \log \tilde{Y}_{PO,t}, X_t, \nu, \Xi, \mu_\delta, A_\delta \sim W \left( \nu + N_t, (\Xi + SSR_t)^{-1} \right) \]  

(B.11)
\[ \delta^{(q)} \mid \Sigma^{(q)}, \log \tilde{Y}_{PO,t}, X_t, \mu_\delta, A_\delta \sim N \left( \text{vec} (\tilde{\Delta}_t), \Sigma^{(q)} \otimes (X_t'X_t + A_\delta)^{-1} \right). \]  

(B.12)

### B.4 Summary

We summarize this adapted Gibbs sampling procedure below.

| Algorithm 3: Gibbs sampling when \( \rho \neq 0 \) |
|---|
| 1 Set \( \{ \delta^{(0)}, \delta_{CP}^{(0)}, \Sigma^{(0)}, \sigma_{CP}^{2(0)}, \mu_\delta, A_\delta, \nu, \Xi, \mu_{\delta CP}, A_{CP}, \alpha_{CP}, \beta_{CP} \} \). |
| for \( (q = 1, \ldots, Q) \) do |
| 2 Draw \( \{ \log Y_i^{\text{miss},(q)}(1), \log Y_i^{\text{miss},(q)}(0), \log B_{CP,i}^{\text{miss},(q)} \}_{i=1}^{N_t} \) using (8) and (B.2)–(B.6). |
| 3 Construct \( \{ \log \tilde{Y}_i^{(q)}(1), \log \tilde{Y}_i^{(q)}(0), \log \tilde{B}_{CP,i}^{(q)} \}_{i=1}^{N_t} \) according to (14). |
| 4 Draw \( \{ \Sigma^{-1,(q)}, \delta^{(q)}, \sigma^{-2,(q)}_{CP}, \delta_{CP}^{(q)} \} \) according to (15)–(19) and (B.7)–(B.12). |
| end |

### B.5 Simulations

To demonstrate the validity of the more general Gibbs sampling algorithm provided above, we replicate the simulations given in section 6. However, we now set \( \rho = 0.6, \nu = 0, \Xi \) to be the identity matrix and the initial value of \( \rho \) at the beginning of each MCMC to 0. All remaining quantities and details of the procedure are the same as described in section 6. Analogous results to those in Figures 2 and 3 are displayed below.
Figure 4: Comparing BTS to A/B tests: interquartiles over 100 rounds across 1,000 simulations
The results in terms of expected regret and convergence of optimality probabilities are very similar to the ones obtained when $\rho = 0$. In only 39 out of the 1,000 simulations the more general algorithm did not stop before the hundredth round, and it found the correct $ATE$ in 98.13% of simulations in which it did stop early.

Figure 5: Histogram of stopping times across simulations with early stoppage

We also experimented with utilizing Algorithm 2 when the DGP is such that $\rho \neq 0$. While this arguably implies that the model becomes misspecified, we found that not only does the algorithm still correctly identifies the true best arm, with results being qualitatively and quantitatively similar to those displayed in section 6, but it is also faster and simpler to implement than Algorithm 3. Hence, in practice it might be preferable to the practitioner to use Algorithm 2 even when it is not assumed that $\rho \neq 0$. However, given the misspecification it is important to acknowledge that such procedure might interfere with the interpretation of the stopping rule suggested in section 5.6.