Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework

Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye & Rui Song

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/01621459.2022.2027776

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework

Chengchun Shi[a*], Xiaoyu Wang[b*], Shikai Luo[c], Hongtu Zhu[d], Jieping Ye[e], and Rui Song[f]

[a]London School of Economics and Political Science, London, UK; [b]Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; [c]ByteDance, Peking, China; [d]The Univeristy of North Carolina at Chapell Hill, Chapel Hill, NC; [e]University of Michigan, Ann Arbor, MI; [f]North Carolina State University, Raleigh, NC

ABSTRACT

A/B testing, or online experiment is a standard business strategy to compare a new product with an old one in pharmaceutical, technological, and traditional industries. Major challenges arise in online experiments of two-sided marketplace platforms (e.g., Uber) where there is only one unit that receives a sequence of treatments over time. In those experiments, the treatment at a given time impacts current outcome as well as future outcomes. The aim of this article is to introduce a reinforcement learning framework for carrying A/B testing in these experiments, while characterizing the long-term treatment effects. Our proposed testing procedure allows for sequential monitoring and online updating. It is generally applicable to a variety of treatment designs in different industries. In addition, we systematically investigate the theoretical properties (e.g., size and power) of our testing procedure. Finally, we apply our framework to both simulated data and a real-world data example obtained from a technological company to illustrate its advantage over the current practice. A Python implementation of our test is available at *https://github.com/callmespring/CausalRL*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2021
Accepted December 2021

KEYWORDS
A/B testing; Causal inference; Online experiment; Online updating; Reinforcement learning; Sequential testing

1. Introduction

A/B testing, or online experiment is a business strategy to compare a new product with an old one in pharmaceutical, technological, and traditional industries (e.g., google, Amazon, or Facebook). It has become the gold standard to make data-driven decisions on a new service, feature, or product. For example, in web analytics, it is common to compare two variants of the same webpage (denote by A and B) by randomly splitting visitors into A and B and then contrasting metrics of interest (e.g., click-through rate) on each of the splits. There is a growing literature on developing A/B testing methods (see e.g., Kharitonov et al. 2015; Johari et al. 2017; Yang et al. 2017, and the references therein). The key idea of these approaches is to apply causal inference methods to estimating the treatment effect of a new change under the assumption of the stable unit treatment value assumption (SUTVA, Rubin 1980; Imbens and Rubin 2015; Wager and Athey 2018, and the references therein). SUTVA precludes the existence of the *interference* effect such that the response of each subject in the experiment depends only on their own treatment and is independent of others' treatments. Despite its ubiquitousness, however, the standard A/B testing is not directly applicable for causal inference under interference (Zhou et al. 2020).

In this article, we focus on the setting where there is only one unit (or system) in the experiment that receives a sequence of treatments over time. In many applications, the treatment at a given time can impact future outcomes, leading SUTVA being invalid. These studies frequently occur in the two-sided markets (intermediary economic platforms having two distinct user groups that provide each other with network benefits) that involve sequential decision making over time. As an illustration, we consider evaluating the effects of different order dispatching strategies in ride-sharing companies (e.g., Uber) for large-scale fleet management. See our real data analysis in Section 5 for details. These companies form a typical two-sided market that enables efficient interactions between passengers and drivers (Rysman 2009). With the rapid development of smart mobile phones and internet of things, they have substantially transformed the transportation landscape of human beings (Jin et al. 2018). Order dispatching is one of the most critical problems in online ride-sharing platforms to adapt the operation and management strategy to the dynamics in demand and supply. At a given time, an order dispatching strategy not only affects the platform's immediate outcome (e.g., passengers' answer time, drivers' income), but also impacts the spatial distribution of drivers in the future. This in turn affects the platform's future outcome. The no interference assumption is thus, violated.

A fundamental question of interest that we consider here is how to develop valid A/B testing methods in the presence of interference. Solving this fundamental question faces at least three major challenges. (i) The first one lies in establishing causal

CONTACT Chengchun Shi ✉ *C.Shi7@lse.ac.uk* London School of Economics and Political Science, London, UK.

*These authors contribute equally to this paper.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

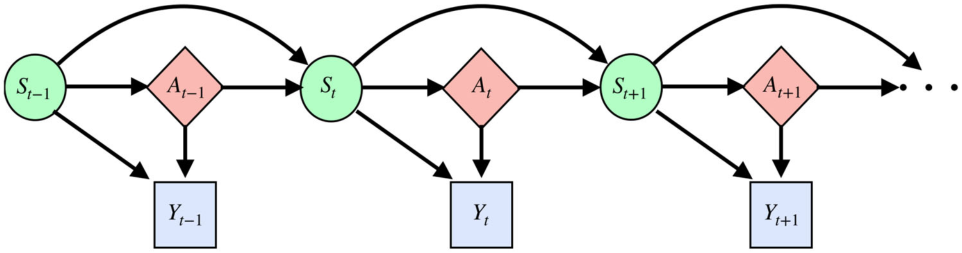


Figure 1. Causal diagram for MDP under settings where treatments depend on current states only. (S_t, A_t, Y_t) represents the state-treatment-outcome triplet. Solid lines represent causal relationships.

relationship between treatments and outcomes over time, by taking the carryover effect into consideration. Most of the existing A/B testing methods are ineffective. They fail to identify the carryover effect, leading the subsequent inference being invalid. See Section 3.1 for details. (ii) The second one is that running each experiment takes a considerable time. The company wishes to terminate the experiment as early as possible in order to save both time and budget. As such, the testing hypothesis needs to be sequentially evaluated online as the data are being collected, and the experiment shall be stopped in accordance with a predefined stopping rule as soon as significant results are observed. (iii) The third one is that treatments are desired to be allocated in a manner to maximize the cumulative outcomes or to detect the alternative more efficiently. The testing procedure shall allow the treatment to be adaptively assigned. Addressing these challenges requires the development of new tools and theory for A/B testing and causal effects evaluation.

1.1. Contributions

We summarize our contributions as follows. First, to address the challenge mentioned in (i), we introduce a reinforcement learning (RL, see, e.g., Sutton and Barto 2018, for an overview) framework for A/B testing. RL is suitable framework to handle the carryover effects over time. In addition to the treatment-outcome pairs, it is assumed that there is a set of time-varying state confounding variables. We model the state-treatment-outcome triplet by using the Markov decision process (MDP, see, e.g., Puterman 1994) to characterize the association between treatments and outcomes across time. Specifically, at each time point, the decision maker selects a treatment based on the observed state variables. The system responds by giving the decision maker a corresponding outcome and moving into a new state in the next time step. In this way, past treatments will have an indirect influence on future rewards through its effect on future state variables. See Figure 1 for an illustration. In addition, the long-term treatment effects can be characterized by the value functions (see Section 2.1 for details) that measure the discounted cumulative gain from a given initial state. Under this framework, it suffices to evaluate the difference between two value functions to compare different treatments. Our proposal gives an example of how to utilize some state-of-the-art machine learning tools, such as reinforcement learning, to address a challenging statistical inference problem for making business decisions.

Second, to address the challenges mentioned in (ii) and (iii), we propose a novel sequential testing procedure for detecting the difference between two value functions. Our proposed test integrates reinforcement learning and sequential analysis (see e.g., Jennison and Turnbull 1999, and the references therein) to allows for sequential monitoring and online updating.[1] Meanwhile, our proposal contributes to each of these two areas as well.

- To the best of our knowledge, this is the first work on developing valid sequential tests in the RL framework. Our work is built upon the temporal-difference learning method based on function approximation (see e.g., Sutton, Szepesvári, and Maei 2008). In the computer science literature, convergence guarantees of temporal difference learning have been derived by Sutton, Szepesvári, and Maei (2008) under the setting of independent noise and by Bhandari, Russo, and Singal (2018) for Markovian noise. However, uncertainty quantification and asymptotic distribution of the resulting value function estimators have been less studied. Such results are critical for carrying out A/B testing. Recently, Luckett et al. (2020) outlined a procedure for estimating the value under a given policy. Shi et al. (2021) developed a confidence interval for the value function. However, these method do not allow for sequential monitoring or online updating.
- Our proposal is built upon the α-spending approach (Lan and DeMets 1983) for sequential testing. We note that most test statistics in classical sequential analysis have the canonical joint distribution (see eq. (3.1) in Jennison and Turnbull 1999) and their associated stopping boundary can be recursively updated via numerical integration. However, in our setup, test statistics no longer have the canonical joint distribution. This is due to the existence of the carryover effects in time. We discuss this in detail in Section 3.4. As such, the numerical integration approach is not applicable to our setting. To resolve this issue, we propose a bootstrap-assisted procedure to determine the stopping boundary. It is much more computationally efficient than the classical wild bootstrap algorithm (Wu 1986, see Section 3.4 for details). The resulting test is generally applicable to a variety of treatment designs, including the Markov design, the alternating-time-interval design and the adaptive design (see Section 3.3 for details).

Third, we systematically investigate the asymptotic properties of our testing procedure. We show that our test not only maintains the nominal Type I error rate, but also has nonneg-

[1] Our test statistic and its stopping boundary are updated as batches of new observations arrive without storing historical data.

ligible powers against local alternatives. In particular, we show that when the sieve method is used for function approximation in temporal difference learning, undersmoothing is not needed to guarantee that the resulting value estimator has a tractable limiting distribution. This occurs because sieve estimators of conditional expectations are idempotent (Newey, Hsieh, and Robins 1998). It implies that the proposed test will not be overly sensitive to the choice of the number of basis functions. To our knowledge, these results have not been established in the existing RL framework. Please see Section 3.3 for details.

Finally, our proposal addresses an important business question in ride-sharing companies. In particular, our methodology allows the company to evaluate different policies more accurately in the presence of carryover effects. It also allows the company to terminate the online experiment earlier and to evaluate more policies within the same time frame. These policies have the potential to improve drivers' salary and meet more customer requests, providing a more efficient transportation network. Please see Section 5 for details.

1.2. Related Work

There is a huge literature on RL in the computer science community such that various algorithms are proposed for an agent to learn an optimal policy and interact with an environment. Recently, a few methods have been developed in the statistics literature on learning the optimal policy in mobile health applications (Ertefaie 2014; Luckett et al. 2020; Hu et al. 2020; Liao, Qi, and Murphy 2020). In addition, there is a growing literature on adapting reinforcement learning to develop dynamic treatment regimes in precision medicine, to recommend treatment decisions based on individual patients' information (Murphy 2003; Chakraborty, Murphy, and Strecher 2010; Qian and Murphy 2011; Zhao et al. 2012; Zhang et al. 2013; Song et al. 2015; Zhao et al. 2015; Zhu et al. 2017; Zhang et al. 2018; Wang et al. 2018; Shi et al. 2018a, 2018b; Mo, Qi, and Liu 2020; Meng et al. 2020).

Our work is closely related to the literature on off-policy evaluation, whose objective is to estimate the value of a new policy based on data collected by a different policy. Existing literature can be cast into model-based methods, importance sampling (IS)-based and doubly-robust procedures. Model-based methods first fit an MDP model from data and then compute the resulting value function. The estimated value function might suffer from a large bias due to potential misspecification of the model. Popular IS based methods include Thomas, Theocharous, and Ghavamzadeh (2015), Thomas and Brunskill (2016), and Liu et al. (2018). These methods reweight the observed rewards with the density ratio of the target and behavior policies. The value estimate might suffer from a large variance, due to the use of importance sampling. Doubly-robust methods (see, e.g., Jiang and Li 2016; Kallus and Uehara 2019) learn the Q-function as well as the probability density ratio and combine these estimates properly for more robust and efficient value evaluation. However, both IS and doubly-robust methods required the treatment assignment probability (propensity score) to be bounded away from 0 and 1. As such, they are inapplicable to the alternating-time-interval design, which is the treatment allocation strategy in our real data application (see, Section 5 for details).

In addition to the literature on RL, our work is also related to a line of research on causal inference with interference. Most of the works studied the interference effect across different subjects (see e.g., Hudgens and Halloran 2008; Pouget-Abadie et al. 2019; Li et al. 2019; Zhou et al. 2020; Reich et al. 2020). That is, the outcome for one subject depends on the treatment assigned to other subjects as well. To the contrary, our work focuses on the interference effect over time. We also remark that most of the aforementioned methods were primarily motivated by research questions in psychological, environmental and epidemiological studies, so their generalization to infer time dependent causal effects in two-sided markets remains unknown.

Finally, we remark that there is a growing literature on evaluating time-varying causal effects (see e.g., Robins 1986; Sobel and Lindquist 2014; Boruvka et al. 2018; Ning, Ghosal, and Thomas 2019; Rambachan and Shephard 2019; Viviano and Bradic 2019; Bojinov and Shephard 2020). However, none of the above cited works used a RL framework to characterize the treatment effects. In particular, Bojinov and Shephard (2020) proposed to use IS based methods to test the null hypothesis of no (average) temporal causal effects in time series experiments. Their causal estimand is different from ours since they focused on p lag treatment effects, whereas we consider the long-term effects characterized by the value function. Moreover, their method requires the propensity score to be bounded away from 0 and 1, and thus, it is not valid for our applications. In addition, these method do not allow for sequential monitoring.

1.3. Organization of the Article

The rest of the article is organized as follows. In Section 2, we introduce a potential outcome framework to MDP and describe the causal estimand. Our testing procedure is introduced in Section 3. In Section 4, we demonstrate the effectiveness of our test via simulations. In Section 5, we apply the proposed test to a data from an online ride-hailing platform to illustrate its usefulness. Finally, we conclude our article in Section 6.

2. Problem Formulation

2.1. A Potential Outcome Framework for MDP

For simplicity, we assume that there are only two treatments (actions, products), coded as 0 and 1, respectively. For any $t \geq 0$, let $\bar{a}_t = (a_0, a_1, \ldots, a_t)^\top \in \{0, 1\}^{t+1}$ denote a treatment history vector up to time t. Let $\mathbb{S}$ denote the support of state variables and S_0 denote the initial state variable. We assume $\mathbb{S}$ is a compact subset of $\mathbb{R}^d$. For any $(\bar{a}_{t-1}, \bar{a}_t)$, let $S_t^*(\bar{a}_{t-1})$ and $Y_t^*(\bar{a}_t)$ be the counterfactual state and counterfactual outcome, respectively, that would occur at time t had the agent followed the treatment history $\bar{a}_t$. The set of potential outcomes up to time t is given by

$$W_t^*(\bar{a}_t) = \{S_0, Y_0^*(a_0), S_1^*(a_0), \ldots, S_t^*(\bar{a}_{t-1}), Y_t^*(\bar{a}_t)\}.$$

Let $W^* = \cup_{t \geq 0, \bar{a}_t \in \{0,1\}^{t+1}} W_t^*(\bar{a}_t)$ be the set of all potential outcomes.

A deterministic policy π is a time-homogeneous function that maps the space of state variables to the set of available actions. Following π, the agent will assign actions according to π at each time. We use $S_t^*(\pi)$ and $Y_t^*(\pi)$ to denote the associated potential state and outcome that would occur at time t had the agent followed π. The goodness of a policy π is measured by its (state) value function,

$$V(\pi; s) = \sum_{t \geq 0} \gamma^t \mathbb{E}\{Y_t^*(\pi) | S_0 = s\},$$

where $0 < \gamma < 1$ is a discount factor that reflects the tradeoff between immediate and future outcomes. The value function measures the discounted cumulative outcome that the agent would receive had they followed π. Note that our definition of the value function is slightly different from those in the existing literature (see, e.g., Sutton and Barto 2018). Specifically, $V(\pi; s)$ is defined through potential outcomes rather than the observed data.

Similarly, we define the Q function by

$$Q(\pi; a, s) = \sum_{t \geq 0} \gamma^t \mathbb{E}\{Y_t^*(\pi(a)) | S_0 = s\},$$

where $\pi(a)$ denotes a time-varying policy where the initial action equals to a and all other actions are assigned according to π.

The goal of A/B testing is to compare the difference between the two treatments. Toward that end, we focus on two nondynamic (state-agnostic) policies that assign the same treatment at each time point. We remark that this is nontraditional in RL where the goal is to build a policy that depends on the state. For these two nondynamic policies, we use their value functions (denote by $V(1; \cdot)$ and $V(0; \cdot)$) to measure their long-term treatment effects. Meanwhile, our proposed method is equally applicable to the dynamic policy scenario as well. To quantitatively compare the two policies, we introduce the Conditional Average Treatment Effect (CATE) and Average Treatment Effect (ATE) based on their value functions in the following definitions. These two definitions relate RL to causal inference.

Definition 1. Conditional on the initial state $S_0 = s$, CATE is defined by the difference between two value functions, that is, $\text{CATE}(s) = V(1; s) - V(0; s)$.

Definition 2. For a given reference distribution function $\mathbb{G}$ that has a bounded density function on $\mathbb{S}$, ATE is defined by the integrated difference between two value function, that is, $\text{ATE} = \int_s \{V(1; s) - V(0; s)\} \mathbb{G}(ds)$.

The focus of this paper is to test the following hypotheses:

$$H_0 : \tau_0 = \text{ATE} \leq 0 \quad \text{versus} \quad H_1 : \tau_0 = \text{ATE} > 0.$$

When H_0 holds, the new product is no better than the old one on average and is not of practical interest.

2.2. Identifiability of ATE

One of the most important question in causal inference is the identifiability of causal effects. In this section, we present sufficient conditions that guarantee the identifiability of the value function.

We first introduce two conditions that are commonly assumed in multi-stage decision making problems (see e.g., Murphy 2003; Robins 2004; Zhang et al. 2013). We need to use the notation $Z_1 \perp\!\!\!\perp Z_2 | Z_3$ to indicate that Z_1 and Z_2 are independent conditional on Z_3. In practice, with the exception of S_0, the set W^* cannot be observed, whereas at time t, we observe the state-action-outcome triplet (S_t, A_t, Y_t). For any $t \geq 0$, let $\bar{A}_t = (A_0, A_1, \ldots, A_t)^\top$ denote the observed treatment history.

(CA) Consistency assumption: $S_{t+1} = S_{t+1}^*(\bar{A}_t)$ and $Y_t = Y_t^*(\bar{A}_t)$ for all $t \geq 0$.

(SRA) Sequential randomization assumption: $A_t \perp\!\!\!\perp W^* | S_t, \{S_j, A_j, Y_j\}_{0 \leq j < t}$.

The CA requires that the observed state and outcome correspond to the potential state and outcome whose treatments are assigned according to the observed treatment history. It generalizes SUTVA to our setting, allowing the potential outcomes to depend on past treatments. The SRA implies that there are no unmeasured confounders and it automatically holds in online randomized experiments, in which the treatment assignment mechanism is prespecified. In SRA, we allows A_t to depend on the observed data history $S_t, \{S_j, A_j, Y_j\}_{0 \leq j < t}$ and thus, the treatments can be adaptively chosen.

We next introduce two conditions that are unique to the reinforcement learning setting.

(MA) Markov assumption: there exists a Markov transition kernel $\mathcal{P}$ such that for any $t \geq 0$, $\bar{a}_t \in \{0, 1\}^{t+1}$ and $\mathcal{S} \subseteq \mathbb{R}^d$, we have $\Pr\{S_{t+1}^*(\bar{a}_t) \in \mathcal{S} | W_t^*(\bar{a}_t)\} = \mathcal{P}(\mathcal{S}; a_t, S_t^*(\bar{a}_{t-1}))$.

(CMIA) Conditional mean independence assumption: there exists a function r such that for any $t \geq 0, \bar{a}_t \in \{0, 1\}^{t+1}$, we have $\mathbb{E}\{Y_t^*(\bar{a}_t) | S_t^*(\bar{a}_{t-1}), W_{t-1}^*(\bar{a}_{t-1})\} = r(a_t, S_t^*(\bar{a}_{t-1}))$.

We make a few remarks. First, these two conditions are central to the empirical validity of reinforcement learning (RL). Specifically, under these two conditions, one can show that there exists an optimal time-homogenous stationary policy whose value is no worse than any history-dependent policy (Puterman 1994). This observation forms the foundation of most of the existing state-of-the-art RL algorithms.

Second, when CA and SRA hold, it implies that the Markov assumption and the conditional mean independence assumption hold on the observed data as well,

$$\Pr(S_{t+1} \in \mathcal{S} | A_t, S_t, \{S_j, A_j, Y_j\}_{0 \leq j < t}) = \mathcal{P}(\mathcal{S}; A_t, S_t), \quad (1)$$

$$\mathbb{E}(Y_t | A_t, S_t, \{S_j, A_j, Y_j\}_{0 \leq j < t}) = r(A_t, S_t). \quad (2)$$

As such, $\mathcal{P}$ corresponds to the transition function that defines the next state distribution conditional on the current state-action pair and r corresponds to the conditional expectation of the immediate reward as a function of the state-action pair.

Assumption (1) is commonly assumed in the existing reinforcement learning literature (see e.g., Ertefaie 2014; Luckett et al. 2020). It is testable based on the observed data. See the goodness-of-fit test developed by Shi et al. (2020). In practice, to ensure the Markov property is satisfied, we can construct the state by concatenating measurements over multiple decision points till the Markovian property is satisfied.

Assumption (2) implies that past treatments will affect future response only through its impact on the future state variables.

In other words, the state variables shall be chosen to include those that serve as important mediators between past treatments and current outcomes. By Assumption (1), this assumption is automatically satisfied when Y_t is a deterministic function of (S_t, A_t, S_{t+1}) that measures the system's status at time $t + 1$. The latter condition is commonly imposed in the reinforcement learning literature and is stronger than (2).

To conclude this section, we derive a version of Bellman equation for the Q function under the potential outcome framework. Specifically, for $a', a \in \{0, 1\}$, let $Q(a'; a, \cdot)$ denote the Q function where treatment a is assigned at the initial decision point and treatment a' is repeatedly assigned afterwards. By definition, we have $V(a; s) = Q(a; a, s)$ for any (a, s).

Lemma 1. Under MA, CMIA, CA, and SRA, for any $t \geq 0$, $a' \in \{0, 1\}$ and any function $\varphi : \mathbb{S} \times \{0, 1\} \to \mathbb{R}$, we have $\mathbb{E}[\{Q(a'; A_t, S_t) - Y_t - \gamma Q(a'; a', S_{t+1})\}\varphi(S_t, A_t)] = 0$.

Lemma 1 implies that the Q-function is estimable from the observed data. Specifically, an estimating equation can be constructed based on Lemma 1 and the Q-function can be learned by solving this estimating equation. Note that $V(a, s) = Q(a; a, s)$ and τ_0 is completely determined by the value function V. As a result, τ_0 =ATE is identifiable.

Note that the positivity assumption is not needed in Lemma 1. Our procedure can thus, handle the case where treatments are deterministically assigned. This is due to MA and CMIA that assume the system dynamics are invariant across time. To elaborate this, note that the discounted value function is completely determined by the transition kernel $\mathcal{P}$ and the reward function r. These quantities can be consistently estimated under certain conditions, regardless of whether the treatments are deterministically assigned or not. Consequently, the value can be consistently estimated even when the treatment assignments are deterministic. We formally introduce our testing procedure in the next section.

3. Testing Procedure

We first introduce a toy example to illustrate the limitations of existing A/B testing methods. We next present our method and prove its consistency under a variety of different treatment designs.

3.1. Toy Examples

Existing A/B testing methods can only detect short-term treatment effects, but fail to identify any long-term effects. To elaborate this, we introduce two examples below.

Example 1. $S_t = 0.5\varepsilon_t$, $Y_t = S_t + \delta A_t$ for any $t \geq 1$ and $S_0 = 0.5\varepsilon_0$.

Example 2. $S_t = 0.5S_{t-1} + \delta A_t + 0.5\varepsilon_t$, $Y_t = S_t$ for any $t \geq 1$ and $S_0 = 0.5\varepsilon_0$.

In both examples, the random errors $\{\varepsilon_t\}_{t\geq 0}$ follow independent standard normal distributions and the parameter δ describes the degree of treatment effects. When $\delta = 0$, H_0 holds.

Table 1. Powers of t-test, DML-based test and the proposed test under Examples 1 and 2, with $T = 500$, $\delta = 0.1$. $\{A_t\}_t$ follow iid Bernoulli distribution with success probability 0.5.

Example 1			Example 2		
t-test 0.76	DML-based test 1	Our test 0.98	t-test 0.04	DML-based test 0.06	Our test 0.73

Suppose $\delta > 0$. Then H_1 holds. In Example 1, the observations are independent and there are no carryover effects at all. In this case, both the existing A/B tests and the proposed test are able to discriminate H_1 from H_0. In Example 2, however, treatments have delayed effects on the outcomes. Specifically, Y_t does not depend on A_t, but is affected by A_{t-1} through S_t. Existing tests will fail to detect H_1 as the short-term conditional average treatment effects $\mathbb{E}(Y_t|A_t = 1, S_t) - \mathbb{E}(Y_t|A_t = 0, S_t) = 0$ in this example. As an illustration, we conduct a small experiment by assuming the decision is made once at $T = 500$, and report the empirical rejection probability of the classical two-sample t-test that is commonly used in online experiments, a more complicated test based on the double machine learning method (Chernozhukov et al. DML, 2017) that is widely employed for inferring causal effects, and the proposed test. It can be seen the competing methods do not have any power under Example 2.

3.2. An Overview of the Proposal

We present an overview of our proposal in this section. As commented before, we adopt a reinforcement learning framework to address the limitations of existing A/B testing methods and characterize the long-term treatment effects. First, we estimate τ_0 based on a version of temporal difference learning. The idea is to apply basis function approximations to solve an estimating equation derived from Lemma 1. Specifically, let $\mathcal{Q} = \{\Psi^\top(s)\beta_a : \beta_a \in \mathbb{R}^q\}$ be a large linear approximation space for $Q(a; a, s) = V(a, s)$, where $\Psi(\cdot)$ is a vector containing q basis functions on $\mathbb{S}$. The dimension q is allowed to grow with the number of samples T to alleviate the effects of model misspecification. Let us suppose $Q \in \mathcal{Q}$ for a moment. Set the function $\psi(s, a)$ in Lemma 1 to $\Psi(s)\mathbb{I}(a = a')$ for $a' = 0, 1$, there exists some $\boldsymbol{\beta}^* = (\beta_0^{*\top}, \beta_1^{*\top})^\top$ such that

$$\mathbb{E}[\{\Psi^\top(S_t)\beta_a^* - Y_t - \gamma\Psi^\top(S_{t+1})\beta_a^*\}\Psi(S_t)\mathbb{I}(A_t = a)] = 0, \quad \forall a \in \{0, 1\},$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The above equations can be rewritten as $\mathbb{E}(\boldsymbol{\Sigma}_t\boldsymbol{\beta}^*) = \mathbb{E}\boldsymbol{\eta}_t$, where $\boldsymbol{\Sigma}_t$ is a block diagonal matrix given by

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \Psi(S_t)\mathbb{I}(A_t = 0)\{\Psi(S_t) - \gamma\Psi(S_{t+1})\}^\top & \\ & \Psi(S_t)\mathbb{I}(A_t = 1)\{\Psi(S_t) - \gamma\Psi(S_{t+1})\}^\top \end{bmatrix}$$

$$\text{and} \quad \boldsymbol{\eta}_t = \{\Psi(S_t)^\top\mathbb{I}(A_t = 0)Y_t, \Psi(S_t)^\top\mathbb{I}(A_t = 1)Y_t\}^\top.$$

Let $\widehat{\boldsymbol{\Sigma}}(t) = t^{-1}\sum_{j<t}\boldsymbol{\Sigma}_j$ and $\widehat{\boldsymbol{\eta}}(t) = t^{-1}\sum_{j<t}\boldsymbol{\eta}_j$. It follows that $\mathbb{E}\{\widehat{\boldsymbol{\Sigma}}(t)\beta^*\} = \mathbb{E}\{\widehat{\boldsymbol{\eta}}(t)\}$. This motivates us to estimate $\boldsymbol{\beta}^*$ by

$$\widehat{\boldsymbol{\beta}}(t) = \{\widehat{\beta}_0^\top(t), \widehat{\beta}_1^\top(t)\}^\top = \widehat{\boldsymbol{\Sigma}}^{-1}(t)\widehat{\boldsymbol{\eta}}(t).$$

ATE can thus, be estimated by the plug-in estimator $\widehat{\tau}(t) = \int_s \Psi^\top(s)\{\widehat{\beta}_1(t) - \widehat{\beta}_0(t)\}\mathbb{G}(ds)$. We remark that there is no guarantee that $\widehat{\boldsymbol{\Sigma}}(t)$ is always invertible. However, its population limit, $\boldsymbol{\Sigma}(t)$ is invertible for any t (see Lemma 3 in the supplementary materials). Consequently, for sufficiently large t, $\widehat{\boldsymbol{\Sigma}}(t)$ is invertible with large probability. In cases where $\widehat{\boldsymbol{\Sigma}}(t)$ is not invertible, we may add a ridge penalty to compute the resulting estimator. See Appendix D.4 of Shi et al. (2021) for details.

Second, we use $\widehat{\tau}(t)$ to construct our test statistic at time t. Let

$$\boldsymbol{U} = \left\{-\int_{s\in\mathbb{S}} \Psi(s)^\top \mathbb{G}(ds), \int_{s\in\mathbb{S}} \Psi(s)^\top \mathbb{G}(ds)\right\}^\top. \tag{3}$$

It follows that $\widehat{\tau}(t) = \boldsymbol{U}\widehat{\boldsymbol{\beta}}(t)$. We will show that $\sqrt{t}\{\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*\}$ is multivariate normal. This implies that $\sqrt{t}\{\widehat{\tau}(t) - \tau_0\}$ is asymptotically normal. Its variance can be consistently estimated by

$$\widehat{\sigma}^2(t) = \boldsymbol{U}^\top \widehat{\boldsymbol{\Sigma}}^{-1}(t)\widehat{\boldsymbol{\Omega}}(t)\{\widehat{\boldsymbol{\Sigma}}^{-1}(t)\}^\top \boldsymbol{U},$$

as t grows to infinity, where $\widehat{\boldsymbol{\Sigma}}^{-1}(t)\widehat{\boldsymbol{\Omega}}(t)\{\widehat{\boldsymbol{\Sigma}}^{-1}(t)\}^\top$ is the sandwich estimator for the variance of $\sqrt{t}\{\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*\}$, and that

$$\widehat{\boldsymbol{\Omega}}(t) = \frac{1}{t}\sum_{j=0}^{t-1} \left\{\begin{array}{l} \Psi(S_j)(1-A_j)\widehat{\varepsilon}_{j,0} \\ \Psi(S_j)A_j\widehat{\varepsilon}_{j,1} \end{array}\right\} \left\{\begin{array}{l} \Psi(S_j)(1-A_j)\widehat{\varepsilon}_{j,0} \\ \Psi(S_j)A_j\widehat{\varepsilon}_{j,1} \end{array}\right\}^\top,$$

where $\widehat{\varepsilon}_{j,a}$ is the temporal difference error $Y_j + \gamma\Psi^\top(S_{j+1})\widehat{\beta}_a - \Psi^\top(S_j)\widehat{\beta}_a$ whose conditional expectation given $(A_j = a, S_j)$ is zero asymptotically (see Lemma 1). This yields our test statistic $\sqrt{t}\widehat{\tau}(t)/\widehat{\sigma}(t)$, at time t. For a given significance level $\alpha > 0$, we reject H_0 when $\sqrt{t}\widehat{\tau}(t)/\widehat{\sigma}(t) > z_\alpha$, where z_α is the upper α-th quantile of a standard normal distribution.

Third, we integrate the α-spending approach with bootstrap to sequentially implement our test (see Section 3.4). The idea is to generate bootstrap samples that mimic the distribution of our test statistics, to specify the stopping boundary at each interim stage. Suppose that the interim analyses are conducted at time points $T_1 < \cdots < T_K = T$. We focus on the setting where both K and $\{T_k\}_k$ are predetermined, as in our application (see Section 5 for details). To simplify the presentation, for each $1 \le k < K$, we assume $T_k/T \to c_k$ for some constants $0 < c_1 < c_2 < \cdots < c_{K-1} < 1$. To better understand our algorithm, we investigate the limiting distribution of our test statistics at these interim stages in the next section.

Finally, we remark that for simplicity, we use the same Q-function model at each interim stage. This works when $\{T_k\}_k$ are of the same order of magnitude, which is the case in our real data application where $T_1 = T_K/2$. Alternatively, one could allow q to grow with k. The testing procedure can be similarly derived.

3.3. Asymptotic Properties Under Different Treatment Designs

We consider three treatment allocation designs that can be handled by our procedure as follows:

D1. Markov design: $\Pr(A_t = 1|S_t, \{S_j, A_j, Y_j\}_{0\le j<t}) = b^{(0)}(S_t)$ for some function $b^{(0)}(\cdot)$ uniformly bounded away from 0 and 1.

D2. Alternating-time-interval design: $A_{2j} = 0, A_{2j+1} = 1$ for all $j \ge 0$.

D3. Adaptive design: For $T_k \le t < T_{k+1}$ for some $k \ge 0$, $\Pr(A_t = 1|S_t, \{S_j, A_j, Y_j\}_{0\le j<t}) = b^{(k)}(S_t)$ for some $b^{(k)}(\cdot)$ that depends on $\{S_j, A_j, Y_j\}_{0\le j<T_k}$ and is uniformly bounded away from 0 and 1 almost surely. We set $T_0 = 0$.

Here, D2 is a deterministic design and is widely used in industry (see our real data example and this technical report[2]). D1 and D3 are random designs. D1 is commonly assumed in the literature on reinforcement learning (Sutton and Barto 2018). D3 is widely employed in the contextual bandit setting to balance the tradeoff between exploration and exploitation. These three settings cover a variety of scenarios in practice.

In D3, we require $b^{(k)}$ to be strictly bounded between 0 and 1. Suppose an ϵ-greedy policy is used, that is, $b^{(k)}(s) = \epsilon/2 + (1-\epsilon)\widehat{\pi}^{(k)}(s)$, where $\widehat{\pi}^{(k)}$ denotes some estimated optimal policy. It follows that $\epsilon/2 \le b^{(k)}(s) \le 1 - \epsilon/2$ for any s. Such a requirement is automatically satisfied. Meanwhile, other adaptive strategies are equally applicable (see e.g., Zhang et al. 2007; Hu, Zhu, and Hu 2015; Metelkina and Pronzato 2017).

For any behavior policy b in D1–D3, define $S_t^*(\bar{b}_{t-1})$ and $Y_t^*(\bar{b}_t)$ as the potential outcomes at time t, where $\bar{b}_t$ denotes the action history assigned according to b. When b is a random policy as in D1 or D3, definitions of these potential outcomes are more complicated than those under a deterministic policy (see Appendix S3, supplementary materials for details). When b is a stationary policy, it follows from MA that $\{S_{t+1}^*(\bar{b}_t)\}_{t\ge -1}$ forms a time-homogeneous Markov chain. When b follows the alternating-time-interval design, both $\{S_{2t}^*(\bar{b}_{2t-1})\}_{t\ge 0}$ and $\{S_{2t+1}^*(\bar{b}_{2t})\}_{t\ge 0}$ form time-homogeneous Markov chains.

To study the asymptotic properties of our test, we need to introduce assumptions C1–C3 and move them and their corresponding detailed discussions to Appendix S4, supplementary materials. In C1, we require the above mentioned Markov chains to be geometrically ergodic. Geometric ergodicity is weaker than the uniform ergodicity condition imposed in the existing reinforcement learning literature (see e.g., Bhandari, Russo, and Singal 2018; Zou, Xu, and Liang 2019). In C2, we impose conditions on the set of basis functions $\Psi(\cdot)$ such that $\Psi(\cdot)$ is chosen to yield a good approximation for the Q function. It is worth mentioning that we only require the approximation error to decay at a rate of $o(T^{-1/4})$ instead of $o(T^{-1/2})$. In other words, "undersmoothing" is not required and the value estimator has a well-tabulated limiting distribution even when the bias of the Q-estimator decays at a rate that is slower than $O(T^{-1/2})$. This result has a number of importation implications. First, it suggests the proposed test will not be overly sensitive to the choice of the number of basis functions. Such a theoretical finding is consistent with our empirical observations in Section 4.3. Second, the number of basis functions could be potentially selected by minimizing the prediction loss of the Q-estimator via cross-validation. We also present examples of basis functions that satisfy C2 in Appendix S4.2, supplementary materials. In C3, we impose some mild conditions on the action

[2] *https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e*

value temporal-difference error, requiring their variances to be nondegenerate.

Let $\{Z_1, \ldots, Z_K\}$ denote the sequence of our test statistics, where $Z_k = \sqrt{T_k}\widehat{\tau}(T_k)/\widehat{\sigma}(T_k)$. In the following, we study their joint asymptotic distributions. We also present an estimator of their covariance matrix that is consistent under all designs.

Theorem 1 (Limiting distributions). Assume C1–C3, MA, CMIA, CA, and SRA hold. Assume all immediate rewards are uniformly bounded variables, the density function of S_0 is uniformly bounded on $\mathbb{S}$ and q satisfies $q = o(\sqrt{T}/\log T)$. Then under D1, D2 or D3, we have

- $\{Z_k\}_{1\le k\le K}$ are jointly asymptotically normal;
- their asymptotic means are nonpositive under H_0;
- their covariance matrix can be consistently estimated by some $\widehat{\boldsymbol{\Xi}}$, whose (k_1, k_2)-th element $\widehat{\Xi}_{k_1,k_2}$ equals

$$\sqrt{\frac{T_{k_1}}{T_{k_2}}}\frac{\boldsymbol{U}^\top \widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_1})\widehat{\boldsymbol{\Omega}}(T_{k_1})\{\widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_2})\}^\top \boldsymbol{U}}{\widehat{\sigma}(T_{k_1})\widehat{\sigma}(T_{k_2})}.$$

This theorem forms the basis of our sequential testing procedure, which we elaborate in the next section.

3.4. Sequential Monitoring and Online Updating

To sequentially monitor our test, we need to specify the stopping boundary $\{b_k\}_{1\le k\le K}$ such that the experiment is terminated and H_0 is rejected when $Z_k > b_k$ for some k.

First, we use the α spending function approach to guarantee the validity of our test. It requires to specify a monotonically increasing function $\alpha(\cdot)$ that satisfies $\alpha(0) = 0$ and $\alpha(T) = \alpha$. Some popular choices of the α spending function include

$$\begin{aligned}&\alpha_1(t) = 2 - 2\Phi\{\Phi^{-1}(1-\alpha/2)\sqrt{T/t}\} \quad \text{and}\\ &\alpha_2(t) = \alpha(t/T)^\theta \quad \text{for } \theta > 0,\end{aligned} \tag{4}$$

where $\Phi(\cdot)$ denotes the normal cumulative distribution function. Adopting the α spending approach, we require b_k's to satisfy

$$\Pr(\cup_{j=1}^k \{Z_j > b_j\}) = \alpha(T_k) + o(1), \qquad \forall 1 \le k \le K. \tag{5}$$

Suppose there exist a sequence of information levels $\{\mathcal{I}_k\}_{1\le k\le K}$ such that

$$\mathrm{cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} + o(1), \tag{6}$$

for all $1 \le k_1 \le k_2$. Then the sequence $\{Z_k\}_{1\le k\le K}$ satisfies the Markov property. The stopping boundary can be efficiently computed based on the numerical integration method detailed in Section 19.2 of Jennison and Turnbull (1999). However, in our setup, condition (6) might not hold when adaptive design is used. As commented in the introduction, this is due to the existence of carryover effects in time. Specifically, when treatment effects are adaptively generated, the behavior policy at difference stages are likely to vary. Due to the carryover effects in time, the state vectors at difference stages have different distribution functions. As such, the asymptotic distribution of the test statistic at each interim stage depends on the behavior policy. Consequently, the covariance $\mathrm{cov}(Z_{k_1}, Z_{k_2})$ is a very complicated function of k_1 and k_2 (see e.g., the form of $\widehat{\Xi}_{k_1,k_2}$ in Theorem 1) that cannot be represented by (6). Consequently, the numerical integration method is not applicable.

Next, we outline a method based on the wild bootstrap (Wu 1986). Then we discuss its limitation and present our proposal, a scalable bootstrap algorithm to determine the stopping boundary. The idea is to generate bootstrap samples $\{\widehat{Z}^{\mathrm{MB}}(t)\}_t$ that have asymptotically the same joint distribution as $\{\sqrt{t}\widehat{\sigma}^{-1}(t)(\widehat{\tau}(t) - \tau_0)\}_t$. By the requirement on $\{b_k\}_k$ in (5), we obtain

$$\begin{aligned}\Pr\left\{Z_k > b_k | \max_{1\le j<k}(Z_j - b_j) \le 0\right\}\\ = \frac{\alpha(T_k) - \alpha(T_{k-1})}{1 - \alpha(T_{k-1})} + o(1).\end{aligned}$$

To implement the test, we thus, recursively calculate the threshold $\widehat{b}_k$ as follows,

$$\begin{aligned}\Pr{}^*\left\{\widehat{Z}^{\mathrm{MB}}(t_k) > \widehat{b}_k | \max_{1\le j<k}(\widehat{Z}^{\mathrm{MB}}(t_j) - \widehat{b}_j) \le 0\right\}\\ = \frac{\alpha(T_k) - \alpha(T_{k-1})}{1 - \alpha(T_{k-1})},\end{aligned} \tag{7}$$

where $\Pr^*$ denotes the probability conditional on the data, and reject H_0 when $Z_k^* > \widehat{b}_k$ for some k. In practice, the above conditional probability can be approximated via Monte Carlo simulations. This forms the basis of the bootstrap algorithm.

Specifically, let $\{\zeta_t\}_{t\ge 0}$ be a sequence of iid mean-zero, unit variance random variables independent of the observed data. Define

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}(t) = \widehat{\boldsymbol{\Sigma}}^{-1}(t)\left[\frac{1}{t}\sum_{j<t}\zeta_j \left\{\begin{array}{c}\Psi(S_j)(1-A_j)\widehat{\varepsilon}_{j,0}\\ \Psi(S_j)A_j\widehat{\varepsilon}_{j,1}\end{array}\right\}\right], \tag{8}$$

where $\widehat{\varepsilon}_{t,a}$ is the temporal difference error defined. Based on $\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}(t)$, one can define the bootstrap sample $\widehat{Z}^{\mathrm{MB}}(t) = \sqrt{t}\widehat{\sigma}^{-1}(t)\boldsymbol{U}^\top\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}(t)$. Based on the definition of $\widehat{\sigma}(t)$, it is immediate to see that each $\widehat{Z}^{\mathrm{MB}}(t)$ follows a standard normal distribution conditional on the data.

We remark that although the wild bootstrap method is developed under the iid settings, it is valid under our setup as well. This is due to that under CMIA, $\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*$ forms a martingale sequence with respect to the filtration $\{(S_j, A_j, Y_j) : j < t\}$. It guarantees that the covariance matrices of $\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}(t)$ and $\widehat{\boldsymbol{\beta}}(t)$ are asymptotically equivalent. As such, the bootstrap approximation is valid.

However, calculating $\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}(T_k)$ requires $O(T_k)$ operations. The time complexity of the resulting bootstrap algorithm is $O(BT_k)$ up to the k-th interim stage, where B is the total number of bootstrap samples. This can be time consuming when $\{T_k - T_{k-1}\}_{k=1}^K$ are large. To facilitate the computation, we observe that in the calculation of $\widehat{\boldsymbol{\beta}}^{\mathrm{MB}}$, the random noise ζ_t is generated upon the arrival of each observation. This is unnecessary as we aim to approximate the distribution of $\widehat{\boldsymbol{\beta}}(\cdot)$ only at finitely many time points $T_1, T_2, \ldots, T_K$.

Finally, we present our bootstrap algorithm to determine $\{b_k\}_{1\le k\le K}$, based on Theorem 1. Let $\{e_k\}_{1\le k\le K}$ be a sequence of iid $N(0, I_{4q})$ random vectors, where I_J stands for a $J \times J$ identity matrix for any J. Let $\widehat{\boldsymbol{\Omega}}(T_0)$ be a zero matrix. At the k-th stage, we compute the bootstrap sample

$$\widehat{Z}_k^* = \frac{\boldsymbol{U}^\top \widehat{\boldsymbol{\Sigma}}^{-1}(T_k)}{\sqrt{T_k}\widehat{\sigma}(T_k)} \sum_{j=1}^{k} \{T_j \widehat{\boldsymbol{\Omega}}(T_j) - T_{j-1}\widehat{\boldsymbol{\Omega}}(T_{j-1})\}^{1/2} e_j.$$

A key observation is that, conditional on the observed dataset, the covariance of $\widehat{Z}_{k_1}^*$ and $\widehat{Z}_{k_2}^*$ equals

$$\frac{\boldsymbol{U}^\top \widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_1})}{\sqrt{T_{k_1} T_{k_2}}\widehat{\sigma}(T_{k_1})\widehat{\sigma}(T_{k_2})} \left[\sum_{j=1}^{k_1} \left\{ T_j \widehat{\boldsymbol{\Omega}}(T_j) - T_{j-1}\widehat{\boldsymbol{\Omega}}(T_{j-1}) \right\} \right] \times \{\widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_2})\}^{-1} \boldsymbol{U} = \widehat{\Xi}_{k_1,k_2}.$$

By Theorem 1, the covariance matrices of $\{Z_k\}_k$ and $\{Z_k^*\}_k$ are asymptotically equivalent. In addition, the limiting distributions of $\{Z_k\}_k$ and $\{Z_k^*\}_k$ are multivariate normal with zero means. As such, the joint distribution of $\{Z_k\}_{1 \le k \le K}$ can be well approximated by that of $\{Z_k^*\}_{1 \le k \le K}$ conditional on the data. The rejection boundary can thus, be computed in a similar fashion as in (7).

Theorem 2 (Type-I error). Suppose that the conditions of Theorem 1 hold and $\alpha(\cdot)$ is continuous. Then the proposed thresholds satisfy $\Pr(\bigcup_{j=1}^k \{Z_j > \widehat{b}_j\}) \le \alpha(T_k) + o(1)$, for all $1 \le k \le K$ under H_0. The equality holds when $\tau_0 = 0$.

Theorem 2 implies that the Type-I error rate of the proposed test is well controlled. When ATE $= 0$, the equality in Theorem 2 holds. The rejection probability achieves the nominal level under H_0. We next investigate the power property of our test.

Theorem 3 (Power). Suppose that the conditions of Theorem 2 hold. Assume $\tau_0 \gg T^{-1/2}$, then $\Pr(\bigcup_{j=1}^k \{Z_j > \widehat{b}_j\}) \to 1$. Assume $\tau_0 = T^{-1/2}h$ for some $h > 0$. Then $\lim_{T \to \infty} [\Pr(\cup_{j=1}^k \{Z_j > \widehat{b}_j\}) - \alpha(T_k)] > 0$.

Combining Theorems 2 and 3 yields the consistency of our test. The second assertion in Theorem 3 implies that our test has nonnegligible powers against local alternatives converging to H_0 at the $T^{-1/2}$ rate. When the signal decays at a slower rate, the power of our test approaches 1.

Since we use a linear basis function to approximate the Q-function, the regression coefficients $\widehat{\boldsymbol{\beta}}(t)$ as well as their covariance estimator can be online updated as batches of observations arrive at the end of each interim stage. As such, our test can be implemented online. We summarize our procedure in Algorithm 1. Recall that q is the number of basis functions. As the kth interim stage, the time complexities of Steps 1–3 in Algorithm 1 are dominated by $O\{q^2(T_k - T_{k-1}) + q^3\}$, $O\{q^2(T_k - T_{k-1}) + q^3\}$ and $O(Bq^2 + q^3)$, respectively. As such, the time complexity of Algorithm 1 is dominated by $O(BKq^2 + Tq^2 + Kq^3)$. In contrast, one can show that the classical wild bootstrap algorithm would take at least $\Omega(BTq^2 + Kq^3)$ number of flops and is much more computationally intensive when $T \gg K$, which is case in phase 3 clinical trials and our real data application.

To conclude this section, we remark that a few bootstrap algorithms have been developed in the RL literature for policy evaluation. Specifically, Hanna, Stone, and Niekum (2017) and Hao et al. (2021) proposed to use bootstrap for uncertainty quantification in off-policy evaluation. These algorithms require the number of trajectories to diverge to infinity to be consistent and are thus, not applicable to our setting where there is only one trajectory in the experiment. In addition, they are developed in offline settings and do not allow online updating. Ramprasad et al. (2021) developed a bootstrap algorithm for policy evaluation in online settings. Their algorithm generates bootstrap samples upon the arrival of each observation and is thus, more computationally intensive than the proposed algorithm.

Algorithm 1 The testing procedure

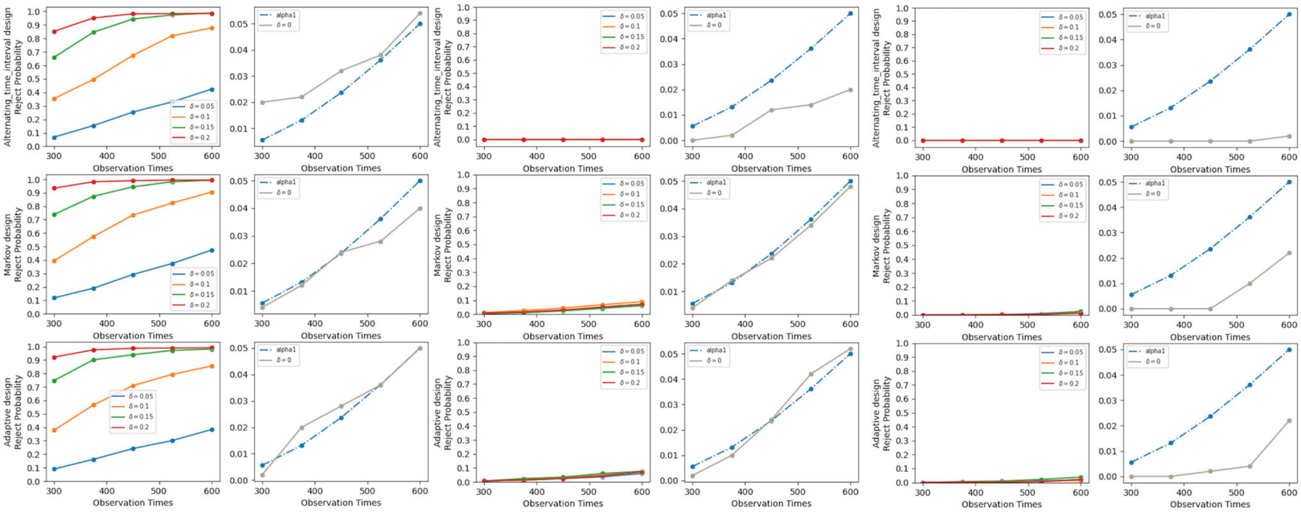
```
Input: number of basis functions $q$, number of bootstrap samples $B$, an $\alpha$ spending function $\alpha(\cdot)$.
Initialize: $T_0 = 0$, $\mathcal{I} = \{1, 2, \ldots, B\}$. Set $\widehat{\boldsymbol{\Omega}}$, $\widehat{\boldsymbol{\Omega}}^*$, $\widehat{\boldsymbol{\Sigma}}_0$, $\widehat{\boldsymbol{\Sigma}}_1$ to zero matrices, and $\widehat{\boldsymbol{\eta}}$, $\widehat{S}_1, \ldots, \widehat{S}_B$ to zero vectors.
Compute $\boldsymbol{U}$ according to (3), using either Monte Carlo methods or numerical integration, where $0_q$ denotes a zero vector of length $q$.
For $k = 1$ to $K$:
    Step 1. Online update of ATE.
    For $t = T_{k-1}$ to $T_k - 1$:
        $\widehat{\boldsymbol{\Sigma}}_a = (1 - t^{-1})\widehat{\boldsymbol{\Sigma}}_a + t^{-1}\Psi(S_t)\mathbb{I}(A_t = a)\{\Psi(S_t) - \gamma\Psi(S_{t+1})\mathbb{I}(A_{t+1} = a)\}^\top$, $a = 0, 1$;
        $\widehat{\boldsymbol{\eta}}_a = (1 - t^{-1})\widehat{\boldsymbol{\eta}}_a + t^{-1}\Psi(S_t)\mathbb{I}(A_t = a)Y_t$.
    Set $\widehat{\beta}_a = \widehat{\boldsymbol{\Sigma}}_a^{-1}\widehat{\boldsymbol{\eta}}_a$ for $a \in \{0, 1\}$ and $\widehat{\tau} = \boldsymbol{U}^\top\widehat{\boldsymbol{\beta}}$.
    Step 2. Online update of the variance estimator.
    Initialize $\widehat{\boldsymbol{\Omega}}^*$ to a zero matrix.
    For $t = T_{k-1}$ to $T_k - 1$:
        $\widehat{\varepsilon}_{t,a} = Y_t + \gamma\Psi^\top(S_{t+1})\widehat{\beta}_a - \Psi^\top(S_t)\widehat{\beta}_a$ for $a = 0, 1$;
        $\widehat{\boldsymbol{\Omega}}^* = \widehat{\boldsymbol{\Omega}}^* + \{\Psi(S_t)^\top(1 - A_t)\widehat{\varepsilon}_{t,0}, \Psi(S_t)^\top A_t\widehat{\varepsilon}_{t,1}\}^\top\{\Psi(S_t)^\top(1 - A_t)\widehat{\varepsilon}_{t,0}, \Psi(S_t)^\top A_t\widehat{\varepsilon}_{t,1}\}$.
    Set $\widehat{\boldsymbol{\Sigma}}$ to a block diagonal matrix by aligning $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_1$ along the diagonal of $\widehat{\boldsymbol{\Sigma}}$;
    Set $\widehat{\boldsymbol{\Omega}} = T_k^{-1}(T_{k-1}\widehat{\boldsymbol{\Omega}} + \widehat{\boldsymbol{\Omega}}^*)$ and the variance estimator $\widehat{\sigma}^2 = \boldsymbol{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Omega}}\{\widehat{\boldsymbol{\Sigma}}^{-1}\}^\top\boldsymbol{U}$.
    Step 3. Bootstrap test statistic.
    For $b = 1$ to $B$:
        Generate $e_k^{(b)} \sim N(0, I_{4q})$;
        $\widehat{S}_b = \widehat{S}_b + \widehat{\boldsymbol{\Omega}}^{*1/2} e_k^{(b)}$;
        $\widehat{Z}_b^* = T_k^{-1/2}\widehat{\sigma}^{-1}\boldsymbol{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{S}_b$;
    Set $z$ to be the upper $\{\alpha(t) - |\mathcal{I}^c|/B\}/(1 - |\mathcal{I}^c|/B)$-th percentile of $\{\widehat{Z}_b^*\}_{b \in \mathcal{I}}$.
    Update $\mathcal{I}$ as $\mathcal{I} \leftarrow \{b \in \mathcal{I} : \widehat{Z}_b^* \le z\}$;
    Step 4. Reject or not?
    Reject the null if $\sqrt{T_k}\widehat{\sigma}^{-1}\widehat{\tau} > z$.
```

4. Simulation Study

4.1. Settings and Implementation

Simulated data of states and rewards was generated as follows,

$$\begin{aligned} S_{1,t} &= (2A_{t-1} - 1)S_{1,(t-1)}/2 + S_{2,(t-1)}/4 + \delta A_{t-1} + \varepsilon_{1,t}, \\ S_{2,t} &= (2A_{t-1} - 1)S_{2,(t-1)}/2 + S_{1,(t-1)}/4 + \delta A_{t-1} + \varepsilon_{2,t}, \\ Y_t &= 1 + (S_{1,t} + S_{2,t})/2 + \varepsilon_{3,t}, \end{aligned}$$



(a) The proposed test under H_1 and H_0

(b) Two-sample t-test under H_1 and H_0

(c) O'Brien & Fleming sequential test under H_1 and H_0

Figure 2. Empirical rejection probabilities of our test and the two-sample t-test with $\alpha(\cdot) = \alpha_1(\cdot)$ and of the O'Brien and Fleming sequential test developed by Kharitonov et al. (2015). The left panels depicts the empirical Type-I error and the right panels depicts the empirical power. Settings correspond to alternating-time-interval, adaptive and Markov designs, from top to bottom plots.

where the random errors $\{\varepsilon_{j,t}\}_{j=1,2,0\le t\le T}$ are iid $N(0, 0.5^2)$ and $\{\varepsilon_{3,t}\}_{0\le t\le T}$ are iid $N(0, 0.3^2)$. Let $S_t = (S_{1,t}, S_{2,t})^\top$ denote the state at time t. Under this model, treatments have delayed effects on the outcomes, as in Example 2. The parameter δ characterizes the degree of such carryover effects. When $\delta = 0$, $\tau_0 = 0$ and H_0 holds. When $\delta > 0$, H_1 holds. Moreover, τ_0 increases as δ increases.

We set $K = 5$ and $(T_1, T_2, T_3, T_4, T_5) = (300, 375, 450, 525, 600)$. The discounted factor γ is set to 0.6 and $\mathbb{G}$ is chosen as the initial state distribution. We consider three behavior policies, according to the designs D1–D3, respectively. For the behavior policy in D1, we set $b^{(0)}(s) = 0.5$ for any $s \in \mathbb{S}$. For the behavior policy in D3, we use an ϵ-greedy policy and set $b^{(k)}(s) = \epsilon/2 + (1-\epsilon)\mathbb{I}(\Psi(s)^\top(\widehat{\beta}_1(T_k) - \widehat{\beta}_0(T_k)) > 0)$, with $\epsilon = 0.1$, for any $k \ge 1$ and $s \in \mathbb{S}$.

For each design, we further consider five choices of δ, corresponding to 0, 0.05, 0.1, 0.15 and 0.2. The significance level α is set to 0.05 in all cases. To implement our test, we choose two α-spending functions, corresponding to $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ given in (4). The hyperparameter θ in $\alpha_2(\cdot)$ is set to 3. The number of bootstrap sample is set to 1000. In addition, we consider the following polynomial basis function, $\Psi(s) = \Psi(s_1, s_2) = (1, s_1, s_1^2, \ldots, s_1^J, s_2, s_2^2, \ldots, s_2^J)^\top$, with $J = 4$.

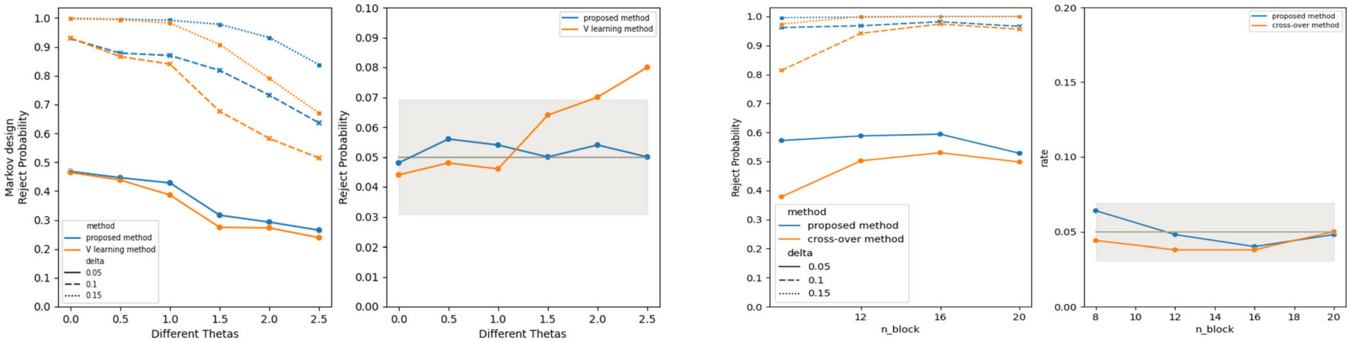
All experiments run on a MacBook Pro with a dual-core 2.7 GHz processor. Implementing a single test takes one second. Figures 2(a) and S1(a) (see Appendix S1, supplementary materials) depict the empirical rejection probabilities of our test statistics at different interim stages under H_0 and H_1 with different combinations of δ, $\alpha(\cdot)$ and the designs. These rejection probabilities are aggregated over 500 simulations. We also plot $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ under H_0. Based on the results, it can be seen that under H_0, the Type-I error rate of our test is well-controlled and close to the nominal level at each interim stage in most cases. Under H_1, the power of our test increases as δ increases, showing the consistency of our test procedure.

4.2. Comparison with Baseline Methods

To further evaluate our method, we first compare it with the classical two-sample t-test and a modified version of modified versions of the O'Brien and Fleming sequential test developed by Kharitonov et al. (2015). We remark that the current practice of policy evaluation in most two-sided marketplace platforms is to employ classical two-sample t-test. Specifically, for each T_k, we apply the t-test to the data $\{A_t, Y_t\}_{0\le t\le T_k}$ and plot the corresponding empirical rejection probabilities in Figures 2(b) and S1(b). Figure 2(c) depicts the empirical rejection probabilities of the modified version of the O'Brien and Fleming sequential test. We remark that such a test requires equal sample size $T_1 = T_k - T_{k-1}$ for $k = 2, \ldots, K$ and is not directly applicable to our setting with unequal sample size. To apply such a test, we modify the decision time and set $(T_1, T_2, T_3, T_4, T_5) = (120, 240, 360, 480, 600)$. As shown in these figures, all these tests fail to detect carryover effects and do not have power at all.

We next compare the proposed test with the test based on the V-learning method developed by Luckett et al. (2020). As we have commented, V-learning does not allow sequential testing. So we focus on settings where the decision is made once at $T = 600$. In addition, V-learning requires the propensity score to be bounded away from 0 and 1. To meet the positivity assumption, we generate the actions according to the Markov design where $\Pr(A_t = 1|S_t) = \text{sigmoid}(\theta S_{1,t} + \theta S_{2,t})$. Both tests require to specify the discounted factor γ. We fix $\gamma = 0.8$. Results are reported in Figure 3(a), aggregated over 500 simulations. It can be seen for large θ, the test based on V-learning cannot control the Type-I error and has smaller power than our test when δ is large. This is because V-learning uses inverse propensity score weighting. In cases where θ is large, the propensity score can be close to zero or one for some sample values, making the resulting test statistic unstable.

Finally, we compare the proposed test with a t-test based on analysis of crossover trials (see e.g., Jones and Kenward 1989).



(a) The proposed test and the test based on V-learning under H_1 and H_0 (from left plots to right plots)

(b) The project test and the t-test derived based on analysis of crossover trials under H_1 and H_0 (from left plots to right plots)

Figure 3. (a) Empirical rejection probabilities of the proposed test and the test based on V-learning. (b) Empirical rejection probabilities of the proposed test and the test derived based on analysis of crossover trials. The shaded area corresponds to the interval $[0.05 - 1.96\text{MCE}, 0.05 + 1.96\text{MCE}]$ where MCE denotes the Monte Carlo error $\sqrt{0.05 \times 0.95/500}$.

We remark that such a test requires the data to be generated from crossover designs and cannot be applied under D1, D2 or D3. In addition, most crossover trials require to recruit multiple subjects/patients to estimate the carryover effect. The resulting tests are not directly applicable to our setting where only one subject receives a sequence of treatments over time. In Appendix S1, supplementary materials, we develop a t-test for the carryover effect under our setting, based on analysis of 2×2 crossover trials. For simplicity, we focus on settings where the decision is made once at $T = 600$. In Figure 3(b), we report the empirical rejection probabilities of such a test and the proposed test under several crossover designs with different number of blocks. Please refer to Appendix S1, supplementary materials for more details about the design and the test. It can be seen that the proposed test is more powerful in most cases.

4.3. *Sensitivity Analysis*

In Section 4.1, we set the number of polynomial basis function J to 4. We also tried some other values of J by setting J to 3 and 5. Results are reported in Figure S2 (see Appendix S1, supplementary materials). It can be seen that the resulting tests have very similar performance and is not sensitive to the choice of J. In Appendix S1, supplementary materials, we fixed J to 4 and tried some other values of $\gamma \in (0.1, 0.3, 0.5, 0.9)$. Results are reported in Figure S3, supplementary materials. It can be seen that our test controls the Type-I error in most cases. In addition, its power increases with γ. This is consistent with the following observation: γ characterizes the balance between the short-term and long-term treatment effects. Under the current setup, there is no short-term treatment effects. The value difference increases with γ. It is thus, expected that our test has better power properties for large values of γ.

5. Real Data Application

We apply the proposed test to a real dataset from a large-scale ride-sharing platform. The purpose of this study is to compare the performance of a newly developed order dispatching strategy with a standard control strategy used in the platform. For a given order, the new strategy will dispatch it to a nearby driver that has not yet finished their previous ride request, but almost. In comparison, the standard control assigns orders to drivers that have completed their ride requests. The new strategy is expected to reduce the chance that the customer will cancel an order in regions with only a few available drivers. It is expected to meet more call orders and increase drivers' income on average.

The experiment is conducted at a given city from December 3 to December 16. Dispatch strategies are executed based on alternating half-hourly time intervals. We also apply our test to a data from an A/A experiment (which compares the baseline strategy against itself), conducted from November 12 to November 25. Note that it is conducted at a different time period from the A/B experiment. The A/A experiment is employed as a sanity check for the validity of the proposed test. We expect that our test will not reject H_0 when applied to this dataset, since the two strategies used are essentially the same.

Both experiments last for two weeks. Thirty-minutes is defined as one time unit. We set $K = 8$ and $T_k = 48 \times (k + 6)$ for $k = 1, \ldots, 8$. That is, the first interim analysis is performed at the end of the first week, followed by seven more at the end of each day during the second week. We discuss more about the experimental design in Appendix S2.2. We choose the overall drivers' income in each time unit as the response. The new strategy is expected to reduce the answer time of passengers and increase drivers' income. Three time-varying variables are used to construct the state. The first two correspond to the number of requests (demand) and drivers' online time (supply) during each 30-minute time interval. These factors are known to have large impact on drivers' income. The last one is the supply and demand equilibrium metric. This variable characterizes the degree that supply meets the demand and serves as an important mediator between past treatments and future outcomes.

To implement our test, we set $\gamma = 0.6$, $B = 1000$ and use a fourth-degree polynomial basis for $\Psi(\cdot)$, as in simulations. We use $\alpha_1(\cdot)$ as the spending function for interim analysis and

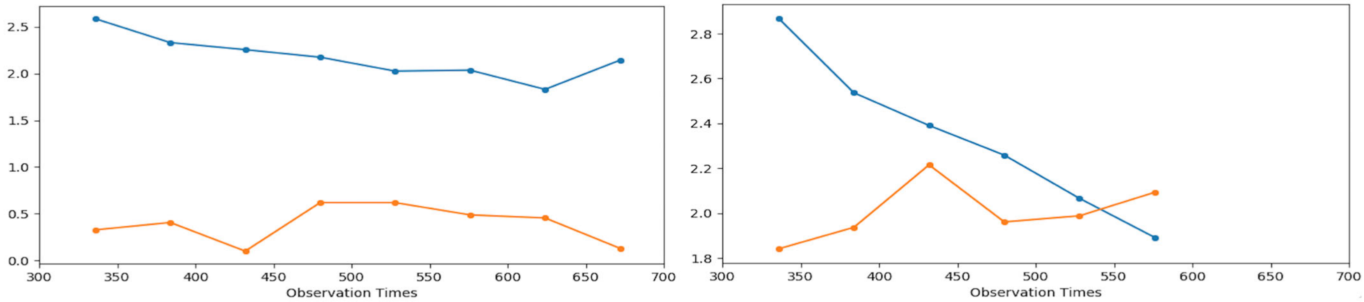


Figure 4. Our test statistic (the orange line) and the rejection boundary (the blue line) in the A/A (left plot) and A/B (right plot) experiments.

set $\alpha = 0.05$. The test statistic and its corresponding rejection boundary at each interim stage are plotted in Figure 4. It can be seen that our test is able to conclude, at the end of the 12th day, that the new order dispatch strategy can significantly increase drivers' income. When applied to the data from the A/A experiment, we fail to reject H_0, as expected. We remark that early termination of the A/B experiment is beneficial to both the platform and the society. First, take this particular experiment as an example, we find that the new strategy reduces the answer time of orders by 2%, leading to almost 2% increment of drivers' income. If we were to wait until Day 14, drivers would lose 2% income and customers would have to wait longer on two days. The benefits are considerable by taking the total number of drivers and customers in the city into account. In addition, the platform can benefit a lot from the increase in the driver income, as they take a fixed proportion of the driving fee from all completed trips. Second, the platform needs to conduct a lot of A/B experiments to investigate various policies. A reduction in the experiment duration facilitates the process, allowing the platform to evaluate more policies within the same time frame. These policies have the potential to further improve the driver income and the customer satisfaction, providing safer, quicker and more convenient transportation.

For comparison, we also apply the two-sample t-test to the data collected from the A/B experiment. The corresponding p-value is 0.18. This result is consistent with our findings. Specifically, the treatment effect at a given time affects the distribution of drivers in the future, inducing interference in time. As shown in the toy example (see Section 3.1), the t-test cannot detect such carryover effects, leading to a low power. Our procedure, according to Theorem 2, has enough powers to discriminate H_1 from H_0.

6. Discussion

First, we remark that our focus is to compare the long-term treatment effects between two nondynamic policies. Meanwhile, the proposed method can be easily extended to handle dynamic policies as well. We discuss this further in Appendix S2.1, supplementary materials.

Second, in our real data application, the design of experiment is determined by the company and we are in the position to analyze the data collected based on such a design. It is important and interesting to design experiments to identity the treatment effect efficiently, but it is beyond the scope of the current paper. In addition, it is worth mentioning that the 30-minute-interval design is adopted by the company to optimize the performance of the resulting A/B test. We consider a few toy examples to elaborate in Appendix S2.2, supplementary materials.

Third, in the current setup, we assume the dimension of the state is fixed whereas the number of basis functions diverges to infinity at a rate that is slower than T. In Appendix S2.3, supplementary materials, we extend our proposal to settings with high-dimensional state information. In that case, we recommend to include a rich class of basis functions to ensure that the Q-function can be well-approximated. The number of basis functions is allowed to be much larger than T. To handle high-dimensionality, we first adopt the Dantzig selector (Candes and Tao 2007) which directly penalizes the Bellman equation to compute an initial estimator. We next develop a decorrelated estimator to reduce the bias of the initial estimator and outline the corresponding testing statistic.

Finally, we focus on causal effects evaluation in online experiments where the treatment generating mechanism is predetermined. Under these settings, there are no unmeasured confounders that confound the action-outcome or the action-next state relationship. Another equally important problem is to study off-policy evaluation in our application. We discuss this further in Appendix S2.4, supplementary materials.

Supplementary Materials

The supplementary materials contain technical assumptions, proofs, additional simulation results, extensions to high-dimensional models and dynamic policies, and discussions on the experimental design, off-policy evaluation and some related works.

Acknowledgments

The authors thank the AE, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article.

Funding

Shi's research was partially supported by the LSE's Research Support Fund in 2021. Song's research was partially supported by grants from NSF-DMS-1555244 and 2113637.

References

Bhandari, J., Russo, D., and Singal, R. (2018), "A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation," arXiv preprint arXiv:1806.02450. [2,6]

Bojinov, I., and Shephard, N. (2020), *Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading*, volume accepted. Taylor & Francis. [3]

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018), “Assessing Time-Varying Causal Effect Moderation in Mobile Health,” *Journal of the American Statistical Association*, 113, 1112–1121. [3]

Candes, E., and Tao, T. (2007), “The Dantzig Selector: Statistical Estimation When p is Much Larger than n,” *Annals of Statistics*, 35, 2313–2351. [11]

Chakraborty, B., Murphy, S., and Strecher, V. (2010), “Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes,” *Statistical Methods in Medical Research*, 19, 317–343. [3]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017), “Double/Debiased/Neyman Machine Learning of Treatment Effects,” *American Economic Review*, 107, 261–65. [5]

Ertefaie, A. (2014), “Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings,” arXiv preprint arXiv:1406.0764. [3,4]

Hanna, J. P., Stone, P., and Niekum, S. (2017), “Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation,” in *Thirty-First AAAI Conference on Artificial Intelligence*. [8]

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. (2021), “Bootstrapping Statistical Inference for Off-Policy Evaluation,” arXiv preprint arXiv:2102.03607. [8]

Hu, J., Zhu, H., and Hu, F. (2015), “A Unified Family of Covariate-Adjusted Response-Adaptive Designs Based on Efficiency and Ethics,” *Journal of the American Statistical Association*, 110, 357–367. [6]

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2020), “Personalized Policy Learning Using Longitudinal Mobile Health Data,” *Journal of the American Statistical Association*, 116, 410–420. [3]

Hudgens, M. G., and Halloran, M. E. (2008), “Toward Causal Inference with Interference,” *Journal of the American Statistical Association*, 103, 832–842. [3]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press. [1]

Jennison, C., and Turnbull, B. W. (1999), *Group Sequential Methods with Applications to Clinical Trials*, Boca Raton, FL: Chapman and Hall/CRC. [2,7]

Jiang, N., and Li, L. (2016), “Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning,” in *International Conference on Machine Learning*, pp. 652–661. [3]

Jin, S. T., Kong, H., Wu, R., and Sui, D. Z. (2018), “Ridesourcing, the Sharing Economy, and the Future of Cities,” *Cities*, 76, 96–104. [1]

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017), “Peeking at a/b Tests: Why it Matters, and What to do About it,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1517–1525. ACM. [1]

Jones, B., and Kenward, M. G. (1989), *Design and Analysis of Cross-Over Trials*, Boca Raton, FL: Chapman and Hall/CRC . [9]

Kallus, N., and Uehara, M. (2019), “Efficiently Breaking the Curse of Horizon: Double Reinforcement Learning in Infinite-Horizon Processes,” arXiv preprint arXiv:1909.05850. [3]

Kharitonov, E., Vorobev, A., Macdonald, C., Serdyukov, P., and Ounis, I. (2015), “Sequential Testing for Early Stopping of Online Experiments,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 473–482. ACM. [1,9]

Lan, K. K. G. and DeMets, D. L. (1983), “Discrete Sequential Boundaries for Clinical Trials,” *Biometrika*, 70, 659–663. [2]

Li, X., Ding, P., Lin, Q., Yang, D., and Liu, J. S. (2019), “Randomization Inference for Peer Effects,” *Journal of the American Statistical Association*, 114, 1651–1664. [3]

Liao, P., Qi, Z., and Murphy, S. (2020), “Batch Policy Learning in Average Reward Markov Decision Processes,” arXiv preprint arXiv:2007.11771. [3]

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018), “Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation,” in *Advances in Neural Information Processing Systems*, pp. 5356–5366. [3]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020), “Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning,” *Journal of the American Statistical Association*, 115, 692–706. [2,3,4,9]

Meng, H., Zhao, Y.-Q., Fu, H., and Qiao, X. (2020), “Near-Optimal Individualized Treatment Recommendations,” arXiv preprint arXiv:2004.02772. [3]

Metelkina, A., and Pronzato, L. (2017), “Information-Regret Compromise in Covariate-Adaptive Treatment Allocation,” *The Annals of Statistics*, 45, 2046–2073. [6]

Mo, W., Qi, Z., and Liu, Y. (2020), “Learning Optimal Distributionally Robust Individualized Treatment Rules,” *Journal of the American Statistical Association*, 116, 659–674. [3]

Murphy, S. A. (2003), “Optimal Dynamic Treatment Regimes,” *Journal of the Royal Statistical Society*, Series B, 65, 331–366. [3,4]

Newey, W. K., Hsieh, F., and Robins, J. (1998), “Undersmoothing and Bias Corrected Functional Estimation,” available at *https://www.researchgate.net/publication/5177172_Undersmoothing_and_Bias_Corrected_Functional_Estimation*. [3]

Ning, B., Ghosal, S., and Thomas, J. (2019), “Bayesian Method for Causal Inference in Spatially-Correlated Multivariate Time Series,” *Bayesian Analysis*, 14, 1–28. [3]

Pouget-Abadie, J., Saint-Jacques, G., Saveski, M., Duan, W., Ghosh, S., Xu, Y., and Airoldi, E. M. (2019), “Testing for Arbitrary Interference on Experimentation Platforms,” *Biometrika*, 106, 929–940. [3]

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: Wiley. [2,4]

Qian, M., and Murphy, S. A. (2011), “Performance Guarantees for Individualized Treatment Rules,” *Annals of Statistics*, 39, 1180–1210. [3]

Rambachan, A., and Shephard, N. (2019), “A Nonparametric Dynamic Causal Model for Macroeconometrics,” *available at SSRN 3345325*. [3]

Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W., and Cheng, G. (2021), “Online Bootstrap Inference for Policy Evaluation in Reinforcement Learning,” arXiv preprint arXiv:2108.03706. [8]

Reich, B. J., Yang, S., Guan, Y., Giffin, A. B., Miller, M. J., and Rappold, A. G. (2020), “A Review of Spatial Causal Inference Methods for Environmental and Epidemiological Applications,” arXiv preprint arXiv:2007.02714. [3]

Robins, J. (1986), “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 7, 1393–1512. [3]

Robins, J. M. (2004), “Optimal Structural Nested Models for Optimal Sequential Decisions,” in *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326. Springer. [4]

Rubin, D. B. (1980), “Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment,” *Journal of the American Statistical Association*, 75, 591–593. [1]

Rysman, M. (2009), “The Economics of Two-Sided Markets,” *Journal of Economic Perspective*, 23, 125–143. [1]

Shi, C., Fan, A., Song, R., and Lu, W. (2018a), “High-Dimensional a-Learning for Optimal Dynamic Treatment Regimes,” *Annals of Statistics*, 46, 925–957. [3]

Shi, C., Song, R., Lu, W., and Fu, B. (2018b), “Maximin Projection Learning for Optimal Treatment Decision with Heterogeneous Individualized Treatment Effects,” *Journal of the Royal Statistical Society*, Series B, 80, 681–702. [3]

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. (2020), “Does the Markov Decision Process Fit the Data: Testing for the Markov Property in Sequential Decision Making,” arXiv preprint arXiv:2002.01751. [4]

Shi, C., Zhang, S., Lu, W., and Song, R. (2021), “Statistical Inference of the Value Function for Reinforcement Learning in Infinite Horizon Settings,” *Journal of the Royal Statistical Society*, Series B, accepted. [2,6]

Sobel, M. E., and Lindquist, M. A. (2014), “Causal Inference for fMRI Time Series Data with Systematic Errors of Measurement in a Balanced On/Off Study of Social Evaluative Threat,” *Journal of the American Statistical Association*, 109, 967–976. [3]

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015), “Penalized q-Learning for Dynamic Treatment Regimens,” *Statistica Sinica*, 25, 901–920. [3]

Sutton, R. S., and Barto, A. G. (2018), *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning (2nd ed.), Cambridge, MA: MIT Press. [2,4,6]

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008), “A Convergent o(n) Algorithm for Off-Policy Temporal-Difference Learning with Linear

Function Approximation," *Advances in Neural Information Processing Systems*, 21, 1609–1616. [2]

Thomas, P., and Brunskill, E. (2016), "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning," in *International Conference on Machine Learning*, pp. 2139–2148. [3]

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. (2015), "High-Confidence Off-Policy Evaluation," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*. [3]

Viviano, D., and Bradic, J. (2019), "Synthetic Learner: Model-Free Inference on Treatments Over Time," arXiv preprint arXiv:1904.01490. [3]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [1]

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018), "Quantile-Optimal Treatment Regimes," *Journal of the American Statistical Association*, 113, 1243–1254. [3]

Wu, C.-F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, 14, 1261–1295. [2,7]

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 *Advances in Neural Information Processing Systems*, pp. 5957–5966. [1]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013), "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694. [3,4]

Zhang, L.-X., Hu, F., Cheung, S. H., Chan, W. S. (2007), "Asymptotic Properties of Covariate-Adjusted Response-Adaptive Designs," *The Annals of Statistics*, 35, 1166–1182. [6]

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018), "Estimation of Optimal Treatment Regimes Using Lists," *Journal of the American Statistical Association*, 113, 1541–1549. [3]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [3]

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [3]

Zhou, Y., Liu, Y., Li, P., and Hu, F. (2020), "Cluster-Adaptive Network a/b Testing: From Randomization to Estimation," arXiv preprint arXiv:2008.08648. [1,3]

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017), "Greedy Outcome Weighted Tree Learning of Optimal Personalized Treatment Rules," *Biometrics*, 73, 391–400. [3]

Zou, S., Xu, T., and Liang, Y. (2019), "Finite-Sample Analysis for Sarsa with Linear Function Approximation," in *Advances in Neural Information Processing Systems*, pp. 8665–8675. [6]