Bias-Variance Tradeoffs for Designing Simultaneous Temporal Experiments

Ruoxuan Xiong ruoxuan.xiong@emory.edu
Emory University

Alex Chin alexchin@lyft.com
Lyft

Sean Taylor seanjtaylor@gmail.com
Motif Analytics

Abstract

We study the analysis and design of simultaneous temporal experiments, where a set of interventions are applied concurrently in continuous time, and outcomes are measured on a sequence of events observed in time. As a motivating setting, suppose multiple data science teams are conducting experiments simultaneously on a ride-hailing platform and independently testing changes to marketplace algorithms such as pricing and matching and estimating effects from observed event outcomes such as the rate at which ride requests are completed. The design problem involves partitioning a continuous space of time into intervals and assigning treatments at the interval level. Design and analysis must account for three factors: carryover effects from interventions at earlier times, correlation in event outcomes, and effects of interventions tested simultaneously. We derive estimators for error components in a highly general setting and build intuition and guidance for practitioners via a careful simulation study.

Keywords: Carryover Effects, Simultaneous Intervention, Treatment Effect Estimation, Temporal Experiment, Experimental Design, Ride-hailing Platform

1 Introduction

In many empirical settings, it is useful to estimate the effects of interventions via time-based or *temporal* experimental designs rather than (the far more common) cross-sectional designs. Most prominently, heuristic designs colloquially known as "switchbacks" have become popular due to their applications in digital marketplaces. In these modern settings, the interference structure between units is difficult to account for and can cause bias of unknown signs and large magnitude using more traditional approaches. Prior to more recent applications, there is a long history in medicine of designing an experiment using a single unit of observation and leveraging longitudinal observations in medicine where it is known as an "n-of-1" trial (Mirza et al., 2017).

As motivation for the present work, we consider the problem of designing multiple simultaneous temporal experiments, for instance, in a ride-hailing company where multiple teams would like to measure the effects of their product changes with only a small number of available treatment units (e.g., cities or regions). In a dynamic two-sided marketplace, users exposed to new pricing and matching algorithms may change their behavior in ways that affect outcomes for other users on either side of the marketplace. There are a variety of causal mechanisms for these spillovers, such as riders consuming available drivers, relocating drivers, or stimulating drivers to drive for longer or shorter periods of time (Chamandy, 2016).

Given the importance of digital marketplaces and the well-acknowledged need to rapidly test new ideas, the design of experiments that provide reliable estimates in the presence of marketplace-mediated interference has drawn increasing attention in recent studies (Holtz et al., 2020; Li et al., 2021; Basse and Feller, 2018; Jagadeesan et al., 2020; Johari et al., 2020). A common theme of these approaches is exploiting prior knowledge of the spillover mechanisms, and leveraging this structure to provide alternative analysis procedures or designs.

We study the design of experiments in a highly generic setting where interventions are applied in a continuous temporal space, and outcomes are measured on a sequence of events in this space. Good designs in this setting efficiently partition continuous temporal space into intervals with alternating treatments in anticipation of precisely estimating a quantity we call the global average treatment effects (GATE) of interventions from the observed event outcomes. GATE is an important estimand for decision-making that captures the difference in average outcomes when an intervention is deployed indefinitely (global treatment) versus when the intervention is absent indefinitely (global control).

Our goal is to capture realistic properties of this empirical setting that complicate the design and analysis of temporal experiments. First, we account for carryover effects between treatments and the outcomes of future events. Second, we account for correlation in event outcomes from unobserved (or unmodeled) factors that create nuisance dependence among measurements; outcomes of events close in time can be similar due to weather, traffic, or other external factors. Correlations do not have to be monotonic in the distance between events, as they can display periodic behavior in weekly or daily cycles. Third, we account for the irregular density of observed events, corresponding to the property that there is strong periodicity in interactions with marketplaces. Finally, we consider the presence of simultaneous experiments run by other teams on the same sequence of events, which can confound effect estimates in finite samples.

Figure 1 introduces the causal structure for our empirical setting. An experimental design is an assignment of $W_{\ell,t}$, which are pure parent nodes, affecting the outcome of n events $Y^{(1)}, \cdots, Y^{(n)}$ occurring at time $t_1, \cdots, t_n$. A latent variable U causes nuisance correlations in all events, yielding dependence in the observations.

To fix a ride-hailing example, consider $W_{p,t}$ as the price charged for a request observed at time t and let $Y^{(i)}$ denote whether the i-th rider's session results in a ride request. The instantaneous effect of an intervention on price is intuitive and immediate. Still, if the ride request occurs, then the effect in a future session may not, due to the diminished supply of available drivers, representing the arrows from W_{p,t_1} to $Y^{(2)}$ and $Y^{(3)}$.

The currently widespread approach to temporal experiments uses a switchback design (Bojinov et al., 2020). Common switchback designs partition time into intervals of equal size and randomizes each interval's treatment assignments. However, fixed-duration switchback designs have two drawbacks. First, the length of time intervals needs to be chosen by the practitioner, which we show theoretically in Section 3 and empirically in Section 4 has a large impact on the performance of the design. Shorter periods increase carryovers from previous

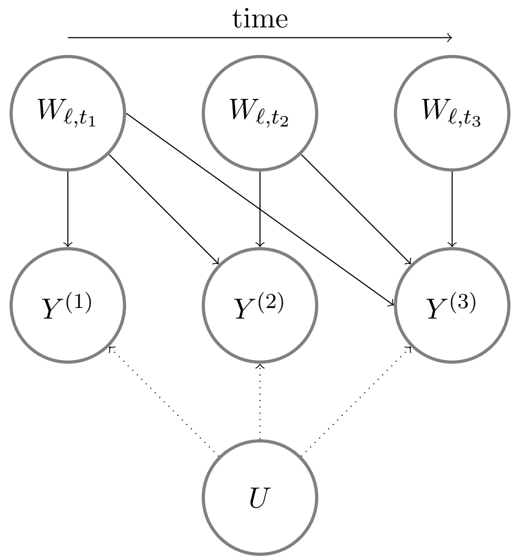


Figure 1: Directed acyclic graph characterizing our empirical setting. W_{ℓ,t_i} denote a vector of interventions applied to time t_i, which affect events $Y^{(i)}$. These instantaneous effects are indicated by solid lines. Past interventions may affect all future events and are observed with nuisance variation (dotted lines) caused by unobserved cause U. The equal spacing of the time of observation is for exposition and is not assumed.

intervals, leading to interference bias. Longer periods decrease precision by decreasing balance in settings with autocorrelation. Second, the absolute time that switching occurs is fixed by choice of period and start time, which limits the ability of the design to exploit information about event density and covariance.

In this paper, we study a setting with multiple interventions and outcomes observed in continuous time. Good designs in this setting will tend to lower the mean square error (MSE) of the estimated GATE estimated using standard Horvitz-Thompson estimators (Horvitz and Thompson, 1952) by effectively trading off various sources of bias and variance. We provide two primary contributions: first, a theoretical analysis to decompose sources of bias and MSE from any design, and second, a simulation study that helps explore these tradeoffs and build intuition for properties of error-minimizing designs.

From a theoretical perspective, in Section 3, we derive a decomposition of both the bias and MSE of the estimated GATE of each intervention. The bias is decomposed into two sources of errors: (a) carryover effects across time of a single intervention, (b) confounding effects from simultaneous interventions. The MSE is caused by both the bias and three sources of randomness: (a) the measurement errors of event outcomes and their covariance, determined by their distance in time, (b) the randomness of treatment assignments at the interval level, (c) the randomness in event occurrence times. The relative contribution of various sources of errors and randomness affects the properties of the optimal treatment design.

To study the temporal experiment design problem empirically, in Section 4, we conduct a simulation study that explores the role of assumptions about carryovers, outcome covariance, and event density in affecting the MSE of heuristic designs. We evaluate the performance of switchbacks with fixed and stochastic periods and characterize the properties of the most efficient designs. Practitioners can use similar simulations with assumptions tailored to their specific design problem in order to design efficient experiments in their empirical settings.

Our results highlight the role of using prior knowledge to select the average period between intervals, the timing of switching, and the role of randomization in improving robustness to carryover effects and simultaneous experiments.

1.1 Related Work

Our work is closely connected to several related literature in the experimental design space. First, there has been extensive work on the design of experiments in temporal or time-series settings, the distinguishing property of which is that outcomes are subject to carryover effects from treatments of prior time periods. As discussed above, the most common tool is the switchback design (Bojinov et al., 2020), in which predetermined time intervals are randomly and sequentially exposed to treatment and control variants. Alternative approaches include pulse designs (Basse and Feller, 2018) where units are treated only for one time period, or designs with irreversible treatment adoption patterns that are based on synthetic control estimators (Doudchenko et al., 2019, 2021; Abadie and Zhao, 2021) or generalized least squares (Xiong et al., 2023).

Designing and analyzing experiments in the presence of interference has been studied in broad settings beyond temporal data. On network data, one common method for mitigating interference is through cluster-randomized designs (Ugander et al., 2013; Eckles et al., 2017; Candogan et al., 2021), where the clusters are chosen to minimize edges that cut across clusters. The cluster size serves an analogous role as the interval length in temporal data, governing the tradeoff between interference bias and estimator variance. Another popular method to mitigate interference is to use two-stage or multi-stage randomization, which has been used in public health (Hudgens and Halloran, 2008; Liu and Hudgens, 2014), political science (Sinclair et al., 2012), and social science (Crépon et al., 2013; Baird et al., 2018; Basse and Feller, 2018). In the spatial setting, a common approach is to conduct experiments at an aggregate level (Bojinov et al., 2020; Xiong et al., 2023) or to randomly assign treatments to a set of predetermined spatial intervention points, with a focus on estimating spatial spillover effects (Aronow et al., 2020, 2021). Our general approach to the temporal problem suggests that some of these ideas may be useful here as well.

Finally, a body of work has been dedicated to the specific setting of marketplace-mediated interference. Johari et al. (2020) study how demand-randomized and supply-randomized designs can contribute different types of bias in a manner that is dependent on market balance. Li et al. (2021) characterize the bias and variance of such experiments and describe how the design can be optimized in such settings. Holtz et al. (2020) compare GATE estimates from a meta-experiment on the Airbnb marketplace that contains both cluster randomization and independent randomization. Holtz and Aral (2020) perform simulation studies that show how cluster-randomized experiments can be effective at reducing bias on the Airbnb network. Our paper takes a more agnostic approach to the marketplace by considering data in the form of a stream of events rather than the explicit two-sided setting.

2 Setting

Suppose K decision makers are simultaneously running experiments on the same time interval. For example, each decision-maker could be on a different team within the same company. Let $T \in \mathbb{R}$ be the experiment duration. The ℓ-th decision maker runs an experiment to study the effect of intervention ℓ, for $\ell \in [K]$, where $[K] = \{1, \cdots, K\}$. For example, the interventions could be pricing, matching, or routing algorithms that are all being tested within the same marketplace in a single region or city. We assume the K interventions are different from one another, and K is finite.

For each intervention $\ell \in [K]$, let $w_{\ell,t} \in \{0, 1\}$ be the treatment status at time $t \in [0, T]$, where $w_{\ell,t} = 1$ indicates that the marketplace is exposed to intervention ℓ (treatment) at time t, and $w_{\ell,t} = 0$ indicates otherwise (control).

Each decision maker ℓ chooses the treatment design of intervention ℓ for the whole experiment horizon, i.e., $\boldsymbol{W}_\ell = \{W_{\ell,t}, \forall t \in [0, T]\}$, pre-experiment. The treatment decisions of all the interventions are made simultaneously. Since the treatment decisions need to be made in a continuous time interval, the decision maker first partitions experimental horizon $[0, T]$ into M disjoint intervals and then randomly chooses the treatment assignment of each interval. For intervention ℓ, let $0 \leq t_{\ell 0} \leq t_{\ell 1} \leq \cdots \leq t_{\ell,M-1} \leq t_{\ell M} = T$ be the endpoints that define the M intervals, let $\mathcal{I}_{\ell m} = [t_{\ell,m-1}, t_{\ell,m}]$ be the m-th interval, and let $|\mathcal{I}_{\ell m}| = t_{\ell m} - t_{\ell,m-1}$ be the length of the m-th interval. For any two interventions, the intervals of one intervention may overlap but not be identical to the intervals of another intervention.[1]

As the treatment decisions are made at the interval level, for all times within an interval, the treatment assignments are the same, i.e.,

$$w_{\ell,t} = w_{\ell,t'}, \qquad \text{for all } t, t' \in \mathcal{I}_{\ell m}, \quad \text{for all } m \text{ and all } \ell.$$

Special cases of these treatment designs include the temporal switchback designs commonly used in practice, where the time intervals are of equal size, and the treatment status is switched randomly between any two consecutive intervals (Chamandy, 2016; Bojinov et al., 2020). Our setup is more general than these commonly used designs in that we allow for intervals of varying length.

The raw data available for analyzing the effect of each intervention are at the event level, where each event could be a rider opening the app and checking the price. Suppose there are n events occurring in the marketplace between time 0 and time T. The outcomes of these n events are available to all decision-makers. Let $Y^{(i)}$ be the outcome of event i that occurred at time t_i, where we assume the occurred time t_i is a random variable. For example, $Y^{(i)}$ could be a binary variable indicating whether the rider requests a ride or not. Let $f(t) : [0, T] \to \mathbb{R}^+$ be the density function from which events are sampled. We assume that $f(t)$ is bounded from below and from above for all t.

1. Without loss of generality, assume M is the same for all interventions by allowing for the interval length to be measure zero.

Example 2.1 (Uniform event density)**.** If events are equally likely to occur at any time in the experiment, then $f(t) = 1/T$ for all $t \in [0, T]$.

We additionally define the marketplace outcome at time t as Y_t. The marketplace outcome Y_t can be viewed as the average outcome of all users in the marketplace, such as the average request rate at time t. The event outcomes are noisy measurements of the marketplace outcomes, i.e., for all i,

$$Y^{(i)} = Y_{t_i} + \varepsilon^{(i)}\,,$$

where the measurement error $\varepsilon^{(i)}$ has mean zero and bounded variance. When the binary $Y^{(i)}$ indicates whether rider i requests a ride, we think of the connection between $Y^{(i)}$ and Y_{t_i} as $Y^{(i)}$ is a random draw from the Bernoulli distribution with probability $\mathbf{P}(Y^{(i)} = 1) = Y_{t_i}$ of being 1. We allow measurement errors of events that are close in time to be correlated:

$$\mathrm{Cov}[\varepsilon^{(i)}, \varepsilon^{(j)}] \neq 0 \qquad \text{for} \ \ t_i \neq t_j\,.$$

The correlation can be caused by external factors like weather, supply conditions, and traffic. This correlation creates a nuisance dependence between event outcomes, which can affect the resulting variance of treatment effect estimates.

Furthermore, we define the potential outcomes of the marketplace at time t as

$$Y_t(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_K)\,,$$

where $\boldsymbol{w}_\ell = \{w_{\ell,t}, \forall t \in [0,T]\}$ is a realization of $\boldsymbol{W}_\ell$.[2] The noisily measured marketplace outcome satisfies $Y_t = Y_t(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K)$. Conditional on treatment designs $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K$ and event occurrence time t_i, there is no randomness in Y_t anymore, and the randomness in $Y^{(i)}$ purely comes from the measurement error $\varepsilon^{(i)}$.

Note that the definition above generalizes the standard, binary definition of potential outcomes under the stable unit treatment value assumption (SUTVA) in two aspects. First, this definition allows potential outcomes to be jointly affected by K interventions. Second, this definition allows for the existence of carryover effects: the potential outcome of t is not only affected by the treatment status at t but also the treatment assignments at other times.

Post-experiment, each decision maker ℓ use observed event outcomes $\{Y^{(i)}\}_{i\in[n]}$ and treatment assignments $\boldsymbol{W}_\ell$ to estimate the effect of intervention ℓ.

2.1 Estimands

Our main object of interest is the *global average treatment effect* (GATE), which measures the difference in average outcomes over time when an intervention is deployed indefinitely (global treatment) versus when an intervention is absent (global control). We formally define the GATE of intervention ℓ as

$$\delta_\ell^{\text{gate}} = \int \delta_{\ell,t}^{\text{gate}} f(t) dt\,,$$

2. Note that intervention ℓ is not applied to times outside of the experiment duration, i.e., $w_{\ell,t}$ is always 0 for $t \notin [0,T]$. Therefore, there are no carryover effects of intervention ℓ from times outside of the experiment duration, $\mathbb{R}\backslash[0,T]$, to the experiment duration, $[0,T]$. It is then reasonable to define potential outcomes only by $\boldsymbol{w}_\ell = \{w_{\ell,t} : \forall t \in [0,T]\}$.

which is the average of total treatment effect $\delta_{\ell,t}^{\text{gate}}$ at time t weighted by the event density $f(t)$. The total treatment effect $\delta_{\ell,t}^{\text{gate}}$ at time t is defined as

$$\delta_{\ell,t}^{\text{gate}} = Y_t(\mathbf{0}, \cdots, \underbrace{\mathbf{1}}_{\boldsymbol{W}_\ell}, \cdots, \mathbf{0}) - Y_t(\mathbf{0}, \cdots, \underbrace{\mathbf{0}}_{\boldsymbol{W}_\ell}, \cdots, \mathbf{0}),$$

where $\mathbf{1} = \{w_{\ell,t} = 1, \forall t \in [0, T]\}$ and $\mathbf{0} = \{w_{\ell,t} = 0, \forall t \in [0, T]\}$ denote global treatment and global control of intervention ℓ, respectively.

We additionally define the average instantaneous effect and average carryover effect, which are building blocks of GATE. The average instantaneous effect $\delta_{\ell}^{\text{inst}}$ is defined as

$$\delta_{\ell}^{\text{inst}} = \int \delta_{\ell,t}^{\text{inst}} f(t) dt,$$

where $\delta_{\ell,t}^{\text{inst}}$ is the instantaneous treatment effect at time t that is defined as

$$\delta_{\ell,t}^{\text{inst}} = Y_t(\mathbf{0}, \cdots, \underbrace{e_t}_{\boldsymbol{W}_\ell}, \cdots, \mathbf{0}) - Y_t(\mathbf{0}, \cdots, \underbrace{\mathbf{0}}_{\boldsymbol{W}_\ell}, \cdots, \mathbf{0})$$

and e_t is a one-hot-encoded vector with the entry of time t to be 1 and all the remaining entries to be 0

$$e_t = \begin{pmatrix} 0 & \cdots & 0 & \underbrace{1}_{\text{time } t} & 0 & \cdots & 0 \end{pmatrix}.$$

The average carryover effect $\delta_{\ell}^{\text{co}}(\boldsymbol{w}_\ell)$, given treatment assignments $\boldsymbol{w}_\ell$, is defined as

$$\delta_{\ell}^{\text{co}}(\boldsymbol{w}_\ell) = \int \delta_{\ell,t}^{\text{co}}(\boldsymbol{w}_\ell) f(t) dt,$$

where $\delta_{\ell,t}^{\text{co}}(\boldsymbol{w}_\ell)$ is the carryover effect at time t that is defined as

$$\delta_{\ell,t}^{\text{co}}(\boldsymbol{w}_\ell) = Y_t(\mathbf{0}, \cdots, \boldsymbol{w}_\ell, \cdots, \mathbf{0}) - Y_t(\mathbf{0}, \cdots, \boldsymbol{w}_\ell \circ e_t, \cdots, \mathbf{0})$$

and "$\circ$" denotes the entry-wise product. Let $\delta_{\ell}^{\text{co}} := \delta_{\ell}^{\text{co}}(\mathbf{1})$ be the average treatment effect under global treatment. Then we can decompose the GATE as

$$\delta_{\ell}^{\text{gate}} = \delta_{\ell}^{\text{inst}} + \delta_{\ell}^{\text{co}}.$$

2.2 Post-Experiment Estimation

Post-experiment, decision makers estimate the GATE using observed event outcomes and treatment designs and decide whether to deploy the intervention indefinitely. We propose to use the Horvitz-Thompson (HT) estimator for $\delta_{\ell}^{\text{gate}}$ (Horvitz and Thompson, 1952), and we analyze the statistical properties of the HT estimator in Section 3.

$$\hat{\delta}_{\ell}^{\text{gate}} = \frac{1}{n} \sum_i \left(\frac{W_{\ell,t_i}}{\pi_\ell} - \frac{1 - W_{\ell,t_i}}{1 - \pi_\ell} \right) Y^{(i)} = \frac{1}{n} \sum_i \alpha_{\ell,t_i} Y^{(i)}, \tag{1}$$

where $\alpha_{\ell,t_i} = \frac{W_{\ell,t_i}-\pi_\ell}{\pi_\ell(1-\pi_\ell)}$ is a normalized weight, and

$$\pi_\ell = \int_{t\in[0,T]} \mathbf{E}[W_{\ell,t}]f(t)dt$$

is the fraction of treated times under intervention ℓ.

We use the HT estimator for three reasons. First, it does not rely on an assumption about carryover mechanisms. Second, it does not rely on assumptions about how the outcomes are correlated in time. Third, it does not require the knowledge of treatment assignments of simultaneous interventions. Due to these three reasons, the HT estimator is flexible and broadly applicable to a wide range of settings in practice.

However, the flexibility of this estimator comes at a cost. First, the HT estimator could be biased due to the carryover effect of the same treatment at other times. The HT estimator approximates the outcomes under global treatment by the event outcomes in treated intervals and approximates the outcomes under global control by the event outcomes in control intervals. When the carryover effect is zero, i.e., $\delta^{\text{co}}_{\ell,t}(\boldsymbol{w}_\ell) = 0$, the approximation error is zero. For general cases, the approximation error is non-zero, and the HT estimator is biased. The bias scales with the size of the carryover effect. Second, the HT estimator can have a large variance as the effective sample size is affected by the correlation of the marketplace (and event) outcomes at different times, and the HT estimator does not optimally weight observations. Third, the HT estimator could have a confounding bias from simultaneous interventions when the treatment designs of two interventions are not orthogonal in finite samples.

It is possible to lower the estimation error of GATE from two aspects. First, we can use a better treatment design, where we provide some guidance in Sections 3 and 4 below. Specifically, in Section 3, we derive a bias-variance decomposition of the estimation error that shows how different sources of errors trade-off. In Section 4, we conduct a simulation study to show how the estimation errors of heuristic designs vary with the assumptions on carryovers, outcome covariance, and event density.

Second, we can use a better estimator for GATE by leveraging prior knowledge of carryover and correlation mechanisms and the information of other interventions. Specifically, to reduce the carryover bias, we can specify the structure of carryover mechanisms, estimate instantaneous and carryover effects, and use them to estimate GATE. To reduce the variance from correlated outcomes, we can specify the structure of correlation mechanisms and use the structure to reweight event outcomes. To reduce confounding bias, we can simultaneously estimate the treatment effects of all interventions. To simultaneously reduce bias and variance, we can use the generalized least squares (GLS) estimators that simultaneously estimate the instantaneous and carryover effects for all interventions by taking advantage of the inverse error covariance weighting. However, our analysis in Section 3 is based on the HT estimator due to the simplicity of exposition, while the design implications are generally the same as the GLS estimators.

2.3 Design of Temporal Switchback Experiments

Before the experiment starts, each decision maker ℓ chooses the treatment design of intervention ℓ, aiming to lower the estimation error of $\hat{\delta}^{\text{gate}}_\ell$, post-experiment. In other words, the decision maker ℓ chooses interval endpoints $t_{\ell m}$ for $m \in \{1, \cdots, M-1\}$, aiming to

$$\min_{t_{\ell 1}, \cdots, t_{\ell, M-1}} \quad \mathbf{E}_{W,\varepsilon,t}\left[(\hat{\delta}_\ell^{\text{gate}} - \delta_\ell^{\text{gate}})^2\right], \tag{2}$$

where the expectation is taken with respect to the treatment designs of all the interventions $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K$, the measurement errors in event outcomes $\varepsilon^{(1)}, \cdots, \varepsilon^{(n)}$, and the event occurrence times $t_1, \cdots, t_n$. Here we focus on the randomized designs, where each interval is equally likely to be treated or untreated, i.e., $\pi_\ell = \mathbf{P}(W_{\ell,t} = 1) = 1/2$ for all ℓ and t. Below we provide two classes of treatment designs to choose interval endpoints, which are referred to as fixed duration and Poisson duration switchbacks.

Example 2.2 (Fixed duration switchbacks)**.** The first interval starts at time $t_{\ell 0} = q$ for some $q < T/M$, and the length of all the intervals beside the last one is T/M. The endpoints then equal to $t_{\ell m} = m \cdot T/M + q$ for all m.

Example 2.3 (Poisson duration switchbacks)**.** The first interval starts at time $t_{\ell 0} = 0$. The length of each interval $t_{\ell m} - t_{\ell,m-1}$ is randomly drawn from the Poisson distribution with the parameter T/M. We sum the lengths of the first to the m-th intervals to obtain the value of the endpoint $t_{\ell m}$.[3]

In Theorem 3.1, we provide the expression of MSE as a function of the interval endpoints, where the interval endpoints can be arbitrarily chosen. The MSE is a complex and nonconvex function of the interval endpoints, so finding the global optimal solution to the optimization problem (2) is generally not feasible. Instead, in Section 4, we provide some general principles for choosing the endpoints for the two classes of designs in Examples 2.2 and 2.3, which can help reduce the objective function value of problem (2).

3 Analysis of Temporal Switchback Designs

In this section, we provide the bias-variance decomposition of the MSE of $\hat{\delta}_\ell^{\text{gate}}$ from the Horvitz-Thompson estimator (1). The decomposition provides insights into how carryovers from interventions at earlier times, correlation in event outcomes, and effects of simultaneous interventions affect the MSE of $\hat{\delta}_\ell^{\text{gate}}$. The insights can then be used as guidance to optimize $\{\boldsymbol{W}_\ell\}_{\ell \in [K]}$ in practice.

We first lay out the assumptions that are necessary for the identification of treatment effects and bias-variance decomposition in Section 3.1. We then introduce several interval-level statistics that measure the average of time-varying components in the potential outcome model over an interval in Section 3.2. Finally, we provide the bias-variance decomposition of the MSE in terms of interval-level statistics in Section 3.3.

3.1 Assumptions

We first assume that the sampling of events is independent of the treatment decisions of all interventions. This assumption makes sense for the interventions that potential riders

3. If the endpoints of some intervals are bigger than the experiment duration (i.e., there exists some $\bar{M}$ such that $t_{\ell m'} > T$ for $m' \geq \bar{M}$), then we set the endpoints of these intervals to T (i.e., set $t_{\ell m'}$ to T for $m' \geq \bar{M}$ and then the lengths of the last $M - \bar{M}$ are zero).

cannot notice a difference before opening the app and checking prices, such as surge pricing algorithms or matching algorithms.

Assumption 3.1 (Exogeneity of events). *Events are sampled randomly and independently from the density function $f(t)$, and $f(t)$ is independent of the treatment assignments of all interventions, $\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_K$.*

Moreover, we assume that carryover effects from the treatments at other times are additive and can be parameterized by a carryover kernel. The additivity assumption is imposed for exposition. It is possible to relax this assumption, at the expense of cumbersome notations in the main results, while the insights are generally the same.

Assumption 3.2 (Carryover effects). *For every t, there exists a non-negative carryover kernel $d^{\mathrm{co}}_{\ell,t}(t')$ that measures the carryover intensity of intervention ℓ from t' to t and satisfies $\int d^{\mathrm{co}}_{\ell,t}(t') f(t') dt' = 1$, such that*

$$\delta^{\mathrm{co}}_{\ell,t}(\boldsymbol{w}_\ell) = \delta^{\mathrm{co}}_{\ell,t} \cdot \int w_{\ell,t'} \cdot d^{\mathrm{co}}_{\ell,t}(t') f(t') dt' .$$

The carryover kernel $d^{\mathrm{co}}_{\ell,t}(t')$ can be quite general in t and t'. Below are four examples of carryover kernels. A visualization of carryover kernels is provided in Figure 2.

Example 3.1 (Uniform carryover kernel). If the carryover intensity is uniform in $t' \in [t-h, t]$, but is zero outside this interval, then $d^{\mathrm{co}}_{\ell,t}(t') \propto 1/h$ for all $t' \in [t-h, t]$ and $d^{\mathrm{co}}_{\ell,t}(t') = 0$ for all $t' \notin [t-h, t]$.

Example 3.2 (Linearly decay carryover kernel). If the carryover intensity decays linearly in $t - t'$ for $t' \in [t-h, t]$, and is zero outside this interval, then $d^{\mathrm{co}}_{\ell,t}(t') \propto t - t'$ for all $t' \in [t-h, t]$ and $d^{\mathrm{co}}_{\ell,t}(t') = 0$ for all $t' \notin [t-h, t]$. See Figure 2a for a visualization.

In addition, we make a simplifying assumption on the structure of intervention effects.

Assumption 3.3 (Additivity of Intervention Effects). *For any intervention ℓ, we have*

$$\begin{aligned} &\mathbf{E}\left[Y_t(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}'_\ell, \cdots, \boldsymbol{w}_K) - Y_t(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_\ell, \cdots, \boldsymbol{w}_K)\right] \\ =&\mathbf{E}\left[Y_t(\boldsymbol{w}'_1, \boldsymbol{w}'_2, \cdots, \boldsymbol{w}'_\ell, \cdots, \boldsymbol{w}'_K) - Y_t(\boldsymbol{w}'_1, \boldsymbol{w}'_2, \cdots, \boldsymbol{w}_\ell, \cdots, \boldsymbol{w}'_K)\right], \end{aligned}$$

where $\boldsymbol{w}_k$ and $\boldsymbol{w}'_k$ are two treatment assignments of intervention k for all k.[4]

When $K = 1$, Assumption 3.3 always holds. When $K > 1$, Assumption 3.3 implies that the effects of K interventions are additive, and this assumption excludes intervention effects to be synergistic (combining two interventions leads to a larger effect than expected) or antagonistic (combining two interventions leads to a smaller effect than expected). Assumption 3.3 is reasonable for certain classes of distinct interventions; for example, we may often assume that a pricing change and a routing change act via different mechanisms and are thus additive. Note that Assumption 3.3 is imposed for the exposition of the MSE decomposition in Section 3.3. It is possible to generalize the MSE decomposition to the case where Assumption 3.3 does not hold.

4. The expected value in Assumption 3.3 is taken with respect to t.

3.2 Interval-Level Statistics

We introduce several interval-level statistics that quantify carryover effects, correlation in measurement errors, confounding effects from simultaneous interventions, and other components at the interval level. These interval-level statistics are building blocks of the bias and mean-squared error decomposition in Section 3.3, and are important quantities to be considered in the partition of intervals.

Fraction of events. Let

$$\mu_\ell^{(m)} = \int_{t_i \in \mathcal{I}_{\ell m}} f(t_i) dt_i$$

be the fraction of events occurring in the interval $\mathcal{I}_{\ell m}$. $\mu_\ell^{(m)}$ ranges from 0 to 1 and the sum of $\mu_\ell^{(m)}$ over m equals to 1.

Example 3.3 (Examples 2.1 and 2.2 (cont.))**.** If event density $f(t)$ is uniform in t and all the intervals are of equal length, then $\mu_\ell^{(m)} = 1/M$.

Mean control outcome. Let

$$\mu_{\ell, Y^{\text{ctrl}}}^{(m)} = \int_{t_i \in \mathcal{I}_{\ell m}} Y_{t_i}(\mathbf{0}, \cdots, \mathbf{0}) f(t_i) dt_i$$

be the integrated global control outcome $Y_{t_i}(\mathbf{0}, \cdots, \mathbf{0})$ over times t_i in the interval $\mathcal{I}_{\ell m}$.

Integrated total treatment effects. Let

$$\Xi_\ell^{(m)} = \int_{t_i \in \mathcal{I}_{\ell m}} \delta_{\ell, t_i}^{\text{gate}} f(t_i) dt_i$$

be the integrated total treatment effect $\delta_{\ell, t_i}^{\text{gate}}$ over times t_i in the interval $\mathcal{I}_{\ell m}$. Following the definition of $\delta_\ell^{\text{gate}}$, the sum of $\Xi_\ell^{(m)}$ over m equals to $\delta_\ell^{\text{gate}}$. Moreover, if treatment effects $\delta_{\ell,t}^{\text{gate}}$ are constant in t, then $\Xi_\ell^{(m)} = \delta_\ell^{\text{gate}} \mu_\ell^{(m)}$ for any m.

Carryover effects. Let

$$I_\ell^{(m,k)} = \int_{t_i \in \mathcal{I}_{\ell m}, t' \in \mathcal{I}_{\ell k}} \delta_{\ell, t_i}^{\text{co}} d_{\ell, t_i}^{\text{co}}(t') f(t_i) f(t') dt_i dt'$$

be integrated carryover effect of treatments at times in the interval $\mathcal{I}_{\ell k}$ on outcomes at times in the interval $\mathcal{I}_{\ell m}$. For notation simplicity, we let $I_\ell^{(m)} = I_\ell^{(m,m)}$ be the integrated carryover effect of treatments on outcomes in the same interval. The integrated carryover effect $I_\ell^{(m,k)}$ increases with the length of both $\mathcal{I}_{\ell m}$ and $\mathcal{I}_{\ell k}$, and increases with the size of carryover effect $\delta_{\ell,t}^{\text{co}}$ for $t \in \mathcal{I}_{\ell m}$. The sum of $I_\ell^{(m,k)}$ over both m and k, which is the integrated carryover effect of the treatment of all intervals on the outcomes of all intervals, is equal to the average carryover effect δ_ℓ^{co}. Moreover, if the carryover effect $\delta_{\ell,t}^{\text{co}}$ is constant in t, then the sum of $I_\ell^{(m,k)}$ over k, which is the integrated carryover effect of the treatment of all intervals on the outcomes in the interval $\mathcal{I}_{\ell m}$, is equal to $\delta_\ell^{\text{co}} \mu_\ell^{(m)}$. Therefore, we can view $I_\ell^{(m,k)}$ as the "building blocks" of δ_ℓ^{co}.

Example 3.4 (Examples 2.1, 2.2, and 3.1 (cont.)). Suppose event density $f(t)$ is uniform in t and all the intervals are of equal length. Furthermore, suppose carryover effect $\delta^{\text{co}}_{\ell,t}$ is constant in t and carryover intensity is constant for $t' \in [t-h, t]$ for any t and for $h < T/M$. Then $I^{(m)}_\ell = \delta^{\text{co}}_\ell(1/M - h/(2T))$.

Variance and covariance of measurement errors. Let the variance of the measurement error of event i occurred at time t_i be (measurement error has mean zero)

$$\text{Var}_{\sigma,t_i} = \mathbf{E}_\varepsilon \left[(\varepsilon^{(i)})^2 \mid t_i\right]$$

and let the corresponding integrated variance of any event that occurred in the interval $\mathcal{I}_{\ell m}$ be

$$V^{(m)}_\ell = \int_{t_i \in \mathcal{I}_{\ell m}} \text{Var}_{\sigma,t_i}\, f(t_i) dt_i\,.$$

Example 3.5 (Examples 2.1 and 2.2 (cont.)). Suppose measurement errors are homoscedastic, i.e., $\text{Var}_{\sigma,i} = \sigma^2$ for all t_i. If the event density $f(t)$ is uniform in t, then $V^{(m)}_\ell = \sigma^2|I_{\ell m}|/T$. Moreover, if all the intervals are of equal length, then $V^{(m)}_\ell = \sigma^2/M$.

Furthermore, let the covariance between the measurement errors of event i occurred at time t_i and event j occurred at time t_j be

$$\text{Cov}_{\sigma,t_i,t_j} = \mathbf{E}_\varepsilon \left[\varepsilon^{(i)}\varepsilon^{(j)} \mid t_i, t_j\right]$$

and let the corresponding integrated covariance between measurement errors of any two events that occurred both in the interval $\mathcal{I}_{\ell m}$ be

$$C^{(m)}_\ell = \int_{t_i,t_j \in \mathcal{I}_{\ell m}} \text{Cov}_{\sigma,t_i,t_j}\, f(t_i) f(t_j) dt_i dt_j\,.$$

When the measurement errors of events at different times are correlated, $C^{(m)}_\ell$ is generally nonzero. In practical settings, there are often some patterns of how the covariance $\text{Cov}_{\sigma,t_i,t_j}$ varies with t_i and t_j, e.g., decays monotonically or periodically in the distance between t_i and t_j. Therefore, we can use a kernel function to parameterize and capture the patterns inherited in $\text{Cov}_{\sigma,t_i,t_j}$. See two examples in Figure 2. Below we show an example of the value of $C^{(m)}_\ell$ when the covariance decays linearly in the distance between t_i and t_j.

Example 3.6 (Example 2.1 (cont.)). Suppose the covariance $\text{Cov}_{\sigma,t_i,t_j}$ decays linearly in $|t_i - t_j|$ for all $t_j \in [t_i - h, t_i + h]$, and is zero outside this interval (i.e., $\text{Cov}_{\sigma,t_i,t_j} = \sigma^2(h - |t_i - t_j|)/h$). Suppose the event density $f(t)$ is uniform in t. If $h < |\mathcal{I}_{\ell m}|$, then $C^{(m)}_\ell = \sigma^2\big(|\mathcal{I}_{\ell m}|^2 - |\mathcal{I}_{\ell m}|h + 2h^2/3\big)/T^2$; otherwise, $C^{(m)}_\ell = \sigma^2\big(|\mathcal{I}_{\ell m}|^2 - |\mathcal{I}_{\ell m}|^3/(3h)\big)/T^2$.

Confounding effects from simultaneous interventions. For intervention ℓ, let

$$S_\ell^{(m)} = \int_{t_i \in \mathcal{I}_{\ell m}} \left[\sum_{\ell' : \ell' \neq \ell} \delta_{\ell', t_i}^{\text{gate}} \right] f(t_i) dt_i$$

be the confounding effects of all the simultaneous interventions on outcomes at times t_i in the interval $\mathcal{I}_{\ell m}$. Furthermore, let

$$S_{\ell,\ell'}^{(m,m')} = \int_{t_i \in \mathcal{I}_{\ell m} \cap \mathcal{I}_{\ell' m'}} \delta_{\ell', t_i}^{\text{inst}} f(t_i) dt_i + \int_{t_j \in \mathcal{I}_{\ell m}, t' \in \mathcal{I}_{\ell' m'}} \delta_{\ell', t_j}^{\text{co}} d_{\ell', t_j}^{\text{co}}(t') f(t') f(t_j) dt_j dt'$$

be the confounding effect of employing intervention ℓ' in the interval $\mathcal{I}_{\ell' m'}$ on outcomes at times t_i in the interval $\mathcal{I}_{\ell m}$. The confounding effect consists of instantaneous and carryover confounding effects. The instantaneous confounding effect only comes from employing intervention ℓ' in the overlapping interval $\mathcal{I}_{\ell m} \cap \mathcal{I}_{\ell' m'}$, while the carryover confounding effect comes from employing intervention ℓ' in the full interval $\mathcal{I}_{\ell' m'}$. Note that if $\mathcal{I}_{\ell m}$ does not overlap with $\mathcal{I}_{\ell' m'}$, then the instantaneous confounding effect is zero.

3.3 Main Results

In this subsection, we show the bias-variance decomposition of the MSE of $\hat{\delta}_\ell^{\text{gate}}$ from the Horvitz-Thompson estimator. The decomposition shows how different components in the potential outcome model affect the estimation error of $\hat{\delta}_\ell^{\text{gate}}$.

Theorem 3.1. *Suppose Assumptions 3.1-3.3 hold and $\pi_\ell = 1/2$. The mean-squared error of $\hat{\delta}_\ell^{\text{gate}}$ equals to*

$$\begin{aligned} \mathbf{E}_{W,\varepsilon,t} \left[\left(\hat{\delta}_\ell^{\text{gate}} - \delta_\ell^{\text{gate}} \right)^2 \right] =& [\text{Bias}_\ell(\text{carryover})]^2 + \text{Var}_\ell(\text{meas}) + \text{Var}_\ell(\text{inst} + \text{carryover}) \\ &+ \text{Var}_\ell(\text{simul}) + 2\,\text{Cov}_\ell(\text{inst} + \text{carryover}, \text{simul}), \end{aligned}$$

where the bias term $\text{Bias}_\ell(\text{carryover})$ *equals to*

$$\text{Bias}_\ell(\text{carryover}) = \sum_{m=1}^M I_\ell^{(m)} - \delta_\ell^{\text{co}};$$

The variance term $\text{Var}_\ell(\text{meas})$ *equals to*

$$\text{Var}_\ell(\text{meas}) = 4 \sum_{m=1}^M \left(V_\ell^{(m)}/n + C_\ell^{(m)} \cdot (n-1)/n \right);$$

The variance term $\text{Var}_\ell(\text{inst} + \text{carryover})$ *equals to*

$$\text{Var}_\ell(\text{inst} + \text{carryover}) = \sum_{m=1}^M \left(\Xi_\ell^{(m)} + 2\mu_{\ell, Y^{\text{ctrl}}}^{(m)} \right)^2 + \sum_{m=1}^M \sum_{m' \neq m} \left(\left[I_\ell^{(m,m')} \right]^2 + I_\ell^{(m,m')} I_\ell^{(m',m)} \right);$$

The variance term $\mathrm{Var}_\ell(\mathrm{simul})$ *equals to*

$$\mathrm{Var}_\ell(\mathrm{simul}) = \sum_{m=1}^{M} \left[S_\ell^{(m)}\right]^2 + \sum_{m=1}^{M}\sum_{m'=1}^{M}\sum_{\ell':\ell'\neq\ell} \left[S_{\ell,\ell'}^{(m,m')}\right]^2 ;$$

The covariance term $\mathrm{Cov}_\ell(\mathrm{inst}+\mathrm{carryover},\mathrm{simul})$ *equals to*

$$\mathrm{Cov}_\ell(\mathrm{inst}+\mathrm{carryover},\mathrm{simul}) = \sum_{m=1}^{M} \left(\Xi_\ell^{(m)} + 2\mu_{\ell,Y^{\mathrm{ctrl}}}^{(m)}\right) S_\ell^{(m)} .$$

Theorem 3.1 shows the bias-variance decomposition of the MSE of $\hat{\delta}_\ell^{\mathrm{gate}}$, using the interval-level statistics defined in Section 3.2. The expectation in Theorem 3.1 is taken with respect to all the randomness in the observations, that is, the treatment designs $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K$, the measurement errors in event outcomes $\varepsilon^{(1)}, \cdots, \varepsilon^{(n)}$, and the event occurrence times $t_1, \cdots, t_n$.

Theorem 3.1 shows that there is only one bias term $\mathrm{Bias}_\ell(\mathrm{carryover})$ in the MSE of $\hat{\delta}_\ell^{\mathrm{gate}}$. This term arises when we use direct treated and control outcomes to approximate globally treated and control outcomes, respectively, in the HT estimator. Continuing Example 3.4 in Section 3.2 where $I_\ell^{(m)} = \delta_\ell^{\mathrm{co}}(1/M - h/(2T))$, the carryover bias equals to $|\,\mathrm{Bias}_\ell(\mathrm{carryover})| = \delta_\ell^{\mathrm{co}} Mh/(2T)$ and increases linearly in the number of intervals M. This implies that switching less frequently can help reduce the bias from carryover effects. See further discussion in Section 4.1 and an example in Figure 3 below.

There are three variance terms in the MSE decomposition of $\hat{\delta}_\ell^{\mathrm{gate}}$. The first variance term $\mathrm{Var}_\ell(\mathrm{meas})$ measures how the event measurement errors affect the MSE of $\hat{\delta}_\ell^{\mathrm{gate}}$. $\mathrm{Var}_\ell(\mathrm{meas})$ consists of two parts: the first part $V_\ell^{(m)}$ measures the variance of measurement error of any event and the second part $C_\ell^{(m)}$ measures the covariance of measurement errors of any two events. As the number of events grows, the first part shrinks to zero and the second part dominates. Following Example 3.6, if intervals have the same length, then $C_\ell^{(m)} = O(1/M^2)$ and $\mathrm{Var}_\ell(\mathrm{error}) = O(1/M)$. This implies that switching more frequently can help reduce the covariance of measurement errors. The intuition that switching more frequently helps is that the number of "interval-level" observations increases, indirectly increasing the effective sample size.

The second variance term $\mathrm{Var}_\ell(\mathrm{inst}+\mathrm{carryover})$ comes from the estimation variance of both instantaneous and carryover effects. The estimation variance of instantaneous effect only contributes to the term $\left(\Xi_\ell^{(m)} + 2\mu_{\ell,Y^{\mathrm{ctrl}}}^{(m)}\right)^2$, while the estimation variance of carryover effect contributes to all the terms. The estimation variance encompasses a trade-off in choosing the number of intervals M. On the one hand, increasing M can increase the variation in treatment assignments at different times, helping to reduce the value of $\sum_{m=1}^{M}\left(\Xi_\ell^{(m)} + 2\mu_{\ell,Y^{\mathrm{ctrl}}}^{(m)}\right)^2$. To see this more clearly, suppose intervals have equal length and both event density $f(t)$ and $\delta_{\ell,t}^{\mathrm{gate}}$ are constant in t. Then $\Xi_\ell^{(m)} = O(1/M)$ and the term $\sum_{m=1}^{M}\left(\Xi_\ell^{(m)} + 2\mu_{\ell,Y^{\mathrm{ctrl}}}^{(m)}\right)^2$ is at the order of $1/M$ and decreases with M. On the other hand, increasing M tends to decrease the length of each interval and increase the carryover effects across intervals, hence increasing the value of $I_\ell^{(m,m')}$ for $m' \neq m$ and decreasing the value of $I_\ell^{(m)}$. Then increasing M increases the value of the other terms in $\mathrm{Var}_\ell(\text{inst}+\text{carryover})$. Therefore, the optimal value of M balances these two competing effects of M on the variance $\mathrm{Var}_\ell(\text{inst}+\text{carryover})$.

Moreover, the choice of interval endpoints also matters for $\mathrm{Var}_\ell(\text{inst}+\text{carryover})$. Consider a special case where $\delta_{\ell,t}^{\text{gate}}$ is constant in t, and $Y_t(\mathbf{0},\cdots,\mathbf{0})$ is constant in t and equals to $\bar{Y}^{\text{ctrl}}$. Then $\sum_{m=1}^{M}\left(\Xi_\ell^{(m)}\right)^2 = (\delta_\ell^{\text{gate}} + 2\bar{Y}^{\text{ctrl}})^2\sum_{m=1}^{M}\left(\mu_\ell^{(m)}\right)^2$, which is minimized at $\mu_\ell^{(m)} = 1/M$ for all m, i.e., equalizing the fraction of events in each interval. This implies that when event density $f(t)$ varies with t, the variance can be lowered by switching more frequently in times of high event density and less frequently in times of low event density.

The third variance term $\mathrm{Var}_\ell(\text{simul})$ comes from confounding effects from simultaneous interventions. This term increases with the scale of confounding effects from simultaneous interventions. In addition, as $S_\ell^{(m)}$ is at the order of $1/M$, the first part of $\mathrm{Var}_\ell(\text{simul})$ is at the order of $1/M$, which can be reduced by increasing M. Moreover, in the second part of $\mathrm{Var}_\ell(\text{simul})$, the term $S_{\ell,\ell'}^{(m,m')}$ is affected by how much $\mathcal{I}_{\ell m}$ overlaps with $\mathcal{I}_{\ell' m'}$. When the interval endpoints of intervention ℓ and ℓ' are the same, $S_{\ell,\ell'}^{(m,m')}$ is the largest. Therefore, misaligning the switching times of different interventions can be helpful to lower the variance.

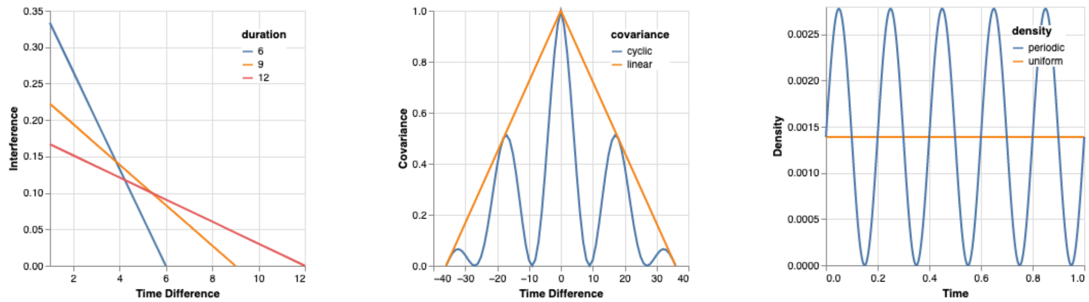
Additionally, there is a covariance term $\mathrm{Cov}_\ell(\text{inst}+\text{carryover},\text{simul})$ in the MSE decomposition of $\hat{\delta}_\ell^{\text{gate}}$, that measures the expected product of simultaneous effects and the sum of instantaneous and carryover effects. To lower this covariance term, it is useful to increase M, following the same reason as the variance terms $\mathrm{Var}_\ell(\text{inst}+\text{carryover})$ and $\mathrm{Var}_\ell(\text{simul})$.

4 Simulation Results

In this section, we present estimates of the MSE of fixed and Poisson duration switchbacks in Examples 2.2 and 2.3 under a simulated problem structure to characterize the tradeoffs involved. Evaluating a design through simulation requires the following inputs:

- Carryover kernel $d_{\ell,t}^{\text{co}}(t')$: we use a linear decay kernel in all simulations.
- Covariance kernel $\mathbf{E}_\varepsilon\left[\varepsilon^{(i)}\varepsilon^{(j)} \mid t_i, t_j\right]$: we consider two regimes, a triangular kernel with height 1, and a periodic covariance kernel which is the product of a triangular kernel and cosine function capturing seasonal patterns.
- Event density $f(t)$: we consider two regimes, uniform density of events and a periodic density $f(t) \propto \sin(\alpha t)$ where events are clustered in time according to a known seasonal pattern.

Figure 2 graphically depicts our design choices for the simulations. Additionally, we vary parameters governing the strength of the instantaneous and carryover effect sizes $\delta_\ell^{\text{inst}}$ and δ_ℓ^{co}, which affect bias from carryover effects. Since these parameters are arbitrary and must be assumed, we choose them such that the resulting bias is on the same scale as the variance.



(a) Carryover kernel (b) Covariance kernel (c) Event density

Figure 2: Simulation setup: carryover and covariance kernels, and event density. Time difference denotes $t'-t$ in the carryover kernel $d^{\text{co}}_{\ell,t}(t')$. If $t'-t<0$, then $d^{\text{co}}_{\ell,t}(t')=0$. The interpretation of time difference is analogous to the covariance kernel.

4.1 Carryover Bias and Variance Tradeoffs

Figure 3 summarizes the most fundamental tradeoff of temporal experiments–policies with shorter periods generate more comparisons that leverage autocorrelation but also increase carryover bias from previous intervals in different conditions. When the carryover effect $\delta^{\text{co}}_{\ell}$ is small, switching as quickly as possible results in the most efficient design, and when it is large, we improve the design by lengthening the period. We focus most of our ensuing discussion on settings where these two error components are on a similar scale and result in an interesting tradeoff.

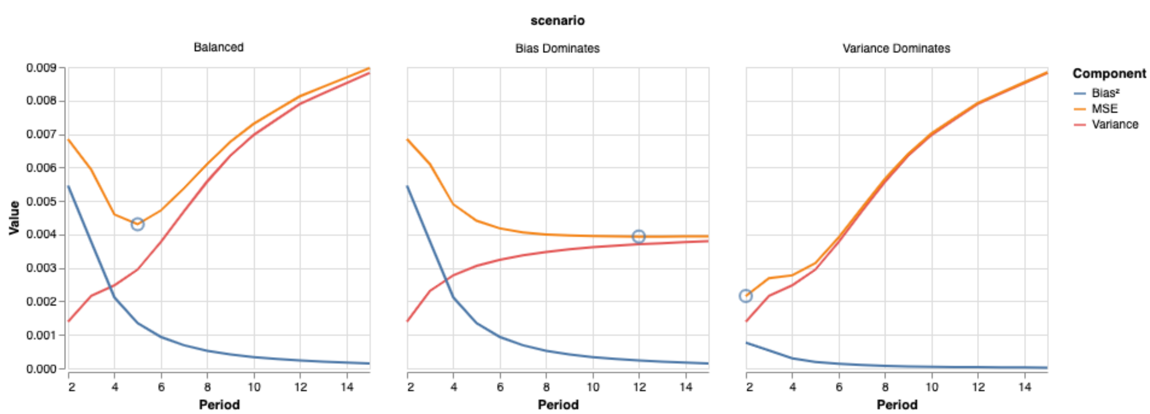


Figure 3: Tradeoffs under different regimes for a deterministic switchback. The x-axis denotes period p in the deterministic switchback with offset $q=0$. The period p with the smallest MSE is circled in blue.

4.2 Stochastic versus Deterministic Designs

Figure 4 compares deterministic designs with various periods to a distribution of errors resulting from different random draws of the stochastic policy. We find that the stochastic switchback generally results in designs with lower bias and increased variance for most values of λ. The randomization generates some longer periods between switching, which helps improve the estimator performance with respect to bias from interference.

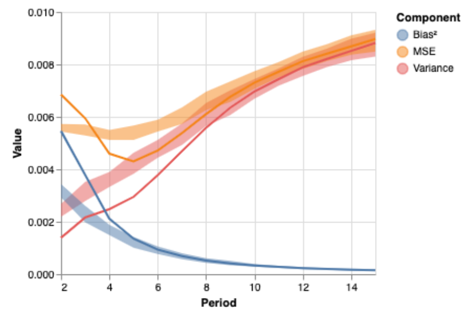


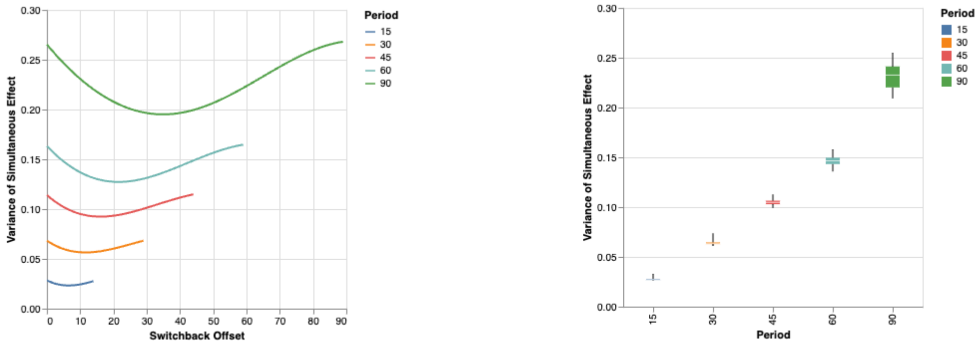
Figure 4: Poisson vs. deterministic switchback. Solid lines denote deterministic switchback. Shaded bands denote Poisson switchback. The x-axis denotes period p in the deterministic switchback and λ in the Poisson switchback.

4.3 Simultaneous Experiments

Our main result in Theorem 3.1 shows that simultaneous experiments increase the estimation variance of the HT estimator. In Figure 5, we show the estimation variance from simultaneous effects when two experiments are run simultaneously. When deterministic switchbacks are used, Figure 5a shows that the estimation variance is affected by both the interval duration and offset in switching times between two experiments. Shortening the interval duration decreases the variance. Moreover, properly staggering two designs also decreases the variance, and the effect is more obvious when the interval duration is longer due to the increased finite-sample correlation between the two designs. Though not depicted, we note that the variance also increases with the number of simultaneous experiments. When Poisson switchbacks are used, Figure 5b shows how the mean period length affects the variance. The Poisson switchback can be more effective unless the deterministic designs are staggered well.

4.4 Periodic Event Density

In many realistic settings, the density of events will exhibit periodic patterns due to the seasonality of human behavior. For instance, in ride-hailing, many ride requests occur during commute times, and relatively few occur during the late evening on weeknights. These daily and weekly cycles create opportunities for improving the design of temporal experiments and motivate simulations with a simple periodic density function. Figure 6 shows results from a periodic density using a deterministic switchback. When the design has a period that aligns with density ($p \in \{6, 12\}$), the offset parameter q determines how



(a) Deterministic switchback

(b) Poisson switchback

Figure 5: Effects of simultaneous experiments. Two simultaneous experiments are run. In Figure 5a, period $p \in \{15, 30, 45, 60, 90\}$ is used in both designs with offset $q = 0$ in one design and varying offset q (x-axis in Figure 5a) in another design. In Figure 5b, we show distributions of variance produced using the Poisson switchback with $\lambda \in \{15, 30, 45, 60, 90\}$ for both designs. The Poisson switchback can be more effective unless the deterministic designs are staggered well.

the alignment alters the bias and variance. For $p = 12$, an offset of 3 (blue dots and lines) yields a design with the lowest variance by switching at an area of maximum density. This results in more events having natural "matches" in an adjacent interval. An offset of 10 (yellow dots and lines) minimizes bias by switching directly after a period with low event density, which minimizes interference from the preceding interval. Knowledge of the density of events can improve the efficiency of the design by leveraging the best absolute times for bias- or variance-minimizing switching points.

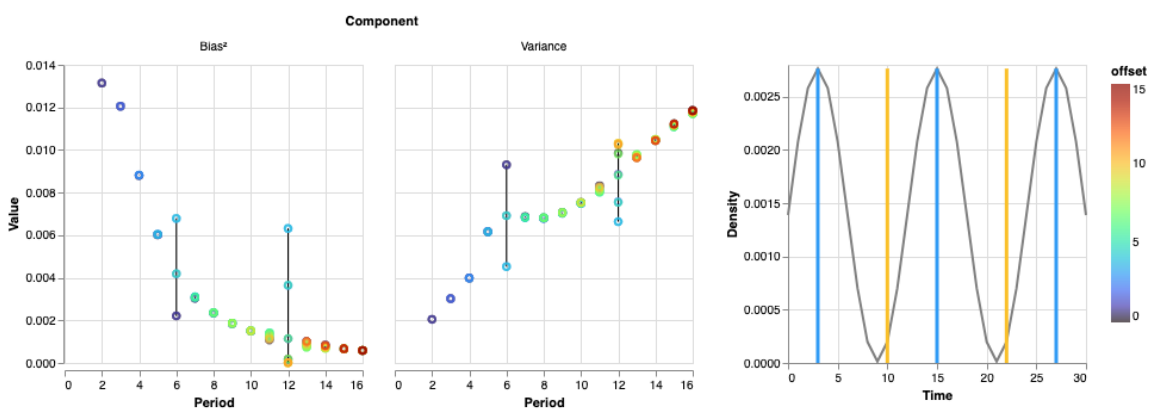


Figure 6: Bias and variance estimates for deterministic switchback in a setting a density with a 12-period cycle. The color of points varies with offset parameter q. In all periods except 6 and 12, the offsets result in almost identical bias and variance.

4.5 Takeaways

Although our simulation results do not allow us to construct an optimal design directly, they point to the properties that better designs would tend to have and the fundamental constraints implied by the noise and causal structure of the setting.

First, as we learned in Section 3, the mean period of the design trades off variance by increasing correlation and bias by decreasing interference from previous periods. We can see that the MSE-minimizing period can vary substantially depending on assumptions about covariance (which are testable), the magnitude of effects, and carryover structure encoded by $d^{\text{co}}_{\ell,t}(t')$, $\delta^{\text{co}}_{\ell}$, and $\delta^{\text{inst}}_{\ell}$ (which must usually be assumed).

Second, stochastic designs exhibit lower bias from interference and simultaneous experiments but do incur some additional variance in order to achieve this. Randomization has the additional benefit that we observe intervals with different lengths, which can help with testing if the treatment causes longer carryovers than were assumed in the design phase.

Third, simultaneous experiments are an important source of error under reasonable assumptions, which is quite a different regime than traditional A/B testing with user-level randomizations, which can generally support many simultaneous tests. In general, the throughput of multiple temporal experiments with substantive effects is something a centralized platform should manage in order to prevent a "tragedy of the commons" result. Ensuring that simultaneous experiments have designs that are uncorrelated in finite samples is likely to be valuable. It could be validated pre-experiment as proposed in Gupta et al. (2018) ("Seedfinder") or restricted randomizations (Simon, 1979).

Fourth, periodic behavior in both event density and in covariance structure implies that there may be benefits and costs to cleverly choosing absolute switching times and periods between switching. A more sophisticated search process could be applied to designing temporal experiments that could leverage estimates of density and the covariance kernel to provide better designs.

5 Discussion and Conclusion

This paper studies the sources of error in the design and analysis of simultaneous temporal experiments. We provide a theoretical analysis of how the bias and variance of the Horvitz-Thompson estimator of the GATE are affected by three factors: carryovers from interventions at earlier times, correlation in event outcomes, and effects of interventions tested concurrently. We provide simulation examples that show how these three factors trade off each other and provide insights into how one can design efficient temporal experiments.

Perhaps the most general conclusion we can draw is that designing experiments in this context involves considering a complex set of tradeoffs and critically depends on the assumptions experimenters would make using prior knowledge. While the expected event density is straightforward to estimate, high-dimensional covariance matrices in event outcomes may pose challenges (Fan et al., 2016). The assumed carryover structure is effectively a causal model that practitioners may need to use prior experimental evidence to capture adequately.

The wide variation in MSE of designs in various simulation setups highlights that useful theory and priors are important factors in the success of experiments in this setting. This is in contrast to randomized experiments with i.i.d. units, where there are various reliable tools for design and analysis, fewer assumptions are needed in either experiment phase, and bias contributes less prominently to estimation (Lin, 2013).

We motivated this study by supporting experiments in a ride-hailing setting where multiple teams share a fixed set of experimental units. Still, we can run experiments over long time periods to increase the sample size. These temporal experiments are a useful tool in this setting, but we could see broader use in other applications with better development of the theory and practical guidelines.

Indeed, there are a variety of settings where cross-sectional interventions are not possible or outcomes cannot be easily attributed to treatment decisions. Estimating the effectiveness of traditional media advertising is well suited to our problem setup, and a privacy-friendly approach to online advertising might employ temporal variation in campaign spending linked to sales through timestamps only. There is also prior work using time-varying interventions in financial or cryptocurrency markets (Krafft et al., 2018) or in self-experimentation for personalized medicine (Karkar et al., 2016). An important goal of this work is to broaden the use of temporal experiments to settings where they are not currently used.

There are important questions left unanswered, and most importantly, a tractable approach for solving the optimal design problem. More sophisticated designs could improve upon the two heuristics we evaluated in Section 4. Solving the globally optimal design that minimizes the MSE is challenging, as conditional on the partitioned time intervals, allocating them to the treatment and control groups is equivalent to the Max-Cut problem that is NP-hard in general graphs. Some heuristic algorithms, such as simulated annealing (Van Laarhoven and Aarts, 1987), or approximation algorithms for Max-Cut, such as randomized rounding (Goemans and Williamson, 1995), could be helpful for finding principled designs and could be another interesting direction to explore for future work.

References

Alberto Abadie and Jinglong Zhao. Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196*, 2021.

Peter M Aronow, Cyrus Samii, and Ye Wang. Design-based inference for spatial experiments with interference. *arXiv preprint arXiv:2010.13599*, 2020.

Peter M Aronow, Cyrus Samii, Jonathan Sullivan, and Ye Wang. Inference in spatial experiments with interference using the spatialeffect package. *arXiv preprint arXiv:2106.15081*, 2021.

Sarah Baird, J Aislinn Bohren, Craig McIntosh, and Berk Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5): 844–860, 2018.

Guillaume Basse and Avi Feller. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020.

Ozan Candogan, Chen Chen, and Rad Niazadeh. Near-optimal experimental design for networks: Independent block randomization. *Available at SSRN*, 2021.

Nicholas Chamandy. Experimentation in a ridesharing marketplace, 2016.

Bruno Crépon, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics*, 128(2):531–580, 2013.

Nick Doudchenko, David Gilinson, Sean Taylor, and Nils Wernerfelt. Designing experiments with synthetic controls. Technical report, Working paper, 2019.

Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, et al. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems*, 34, 2021.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices, 2016.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Somit Gupta, Lucy Ulanova, Sumit Bhardwaj, Pavel Dmitriev, Paul Raff, and Aleksander Fabijan. The anatomy of a large-scale experimentation platform. In *2018 IEEE International Conference on Software Architecture (ICSA)*, pages 1–109. IEEE, 2018.

David Holtz and Sinan Aral. Limiting bias from test-control interference in online marketplace experiments. *Available at SSRN 3583596*, 2020.

David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. Reducing interference bias in online marketplace pricing experiments. *Available at SSRN 3583836*, 2020.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Ravi Jagadeesan, Natesh S Pillai, and Alexander Volfovsky. Designs for estimating the treatment effect in networks with interference. *The Annals of Statistics*, 48(2):679–712, 2020.

Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Weintraub. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670*, 2020.

Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R Mishra, James Fogarty, Sean A Munson, and Julie A Kientz. A framework for self-experimentation in personalized health. *Journal of the American Medical Informatics Association*, 23(3):440–448, 2016.

Peter M Krafft, Nicolás Della Penna, and Alex Sandy Pentland. An experimental study of cryptocurrency market dynamics. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. *arXiv preprint arXiv:2104.12222*, 2021.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Lan Liu and Michael G Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association*, 109(505): 288–301, 2014.

RD Mirza, S Punja, S Vohra, and G Guyatt. The history and development of n-of-1 trials. *Journal of the Royal Society of Medicine*, 110(8):330–340, 2017.

Richard Simon. Restricted randomization designs in clinical trials. *Biometrics*, pages 503–512, 1979.

Betsy Sinclair, Margaret McConnell, and Donald P Green. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56 (4):1055–1069, 2012.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.

Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

Ruoxuan Xiong, Susan Athey, Mohsen Bayati, and Guido W Imbens. Optimal experimental design for staggered rollouts. *Management Science*, 2023.