Optimal Treatment Allocation for Efficient Policy Evaluation in Sequential Decision Making

Ting Li[1] **Chengchun Shi**[2] **Jianing Wang**[1] **Fan Zhou**[1] **Hongtu Zhu**[3] *
[1]School of Statistics and Management, Shanghai University of Finance and Economics
[2]Department of Statistics, London School of Economics and Political Science
[3]Department of Biostatistics, University of North Carolina at Chapel Hill
tingli@mail.shufe.edu.cn, c.shi7@lse.ac.uk, jianing.wang@163.sufe.edu.cn
zhoufan@mail.shufe.edu.cn,htzhu@email.unc.edu

Abstract

A/B testing is critical for modern technological companies to evaluate the effectiveness of newly developed products against standard baselines. This paper studies optimal designs that aim to maximize the amount of information obtained from online experiments to estimate treatment effects accurately. We propose three optimal allocation strategies in a dynamic setting where treatments are sequentially assigned over time. These strategies are designed to minimize the variance of the treatment effect estimator when data follow a non-Markov decision process or a (time-varying) Markov decision process. We further develop estimation procedures based on existing off-policy evaluation (OPE) methods and conduct extensive experiments in various environments to demonstrate the effectiveness of the proposed methodologies. In theory, we prove the optimality of the proposed treatment allocation design and establish upper bounds for the mean squared errors of the resulting treatment effect estimators.

1 Introduction

Motivation. Prior to the full-scale deployment of any product in practical applications, an accurate evaluation of its potential impact is crucial. Modern technology companies, including Amazon, Google, Netflix, Uber, and Didi, commonly employ online experimentation or A/B testing as a means of evaluating the effectiveness of new products or policies (treatment) in comparison to their existing counterparts (control). Often, in these experiments, policies are assigned sequentially over time, impacting both current and future responses. Such a dynamic can invalidate the Stable Unit Treatment Value Assumption (SUTVA), as outlined by (Imbens and Rubin, 2015). Failure to consider these temporal carryover effects can lead to a biased estimation of the treatment effect. Additionally, the limited duration of the experiment and the typically small size of the treatment effect pose significant challenges to consistently detecting the treatment effect.

A substantial body of literature has been dedicated to developing policy evaluation or A/B testing algorithms using data from online experiments. However, there has been limited focus on the generation of the experimental dataset itself. This factor is critical as it can substantially influence the precision of the subsequent treatment effect estimator. This paper's primary focus is to study optimal experimental designs in the context of sequential decision making. In clinical trials, a carefully designed experiment can significantly improve the accuracy of the treatment effect estimator and the statistical power of the associated test, as noted by (Sverdlov et al., 2020). However, most existing

*The first two authors contribute equally to this paper. Address for correspondence: Hongtu Zhu, Ph.D., E-mail: htzhu@email.unc.edu.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

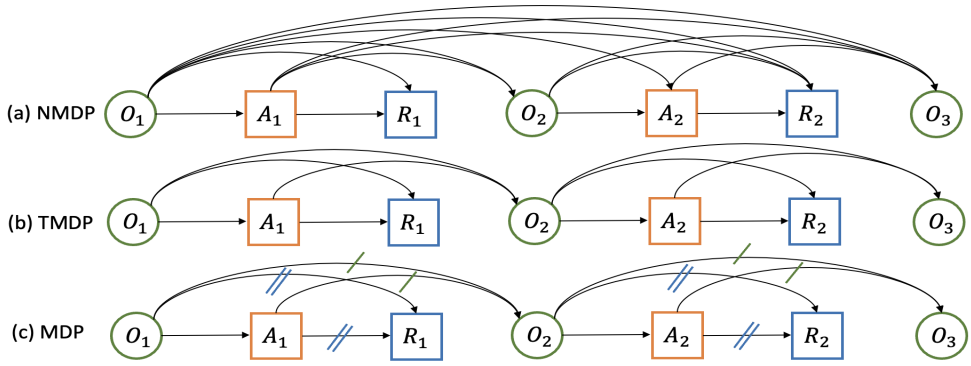

Figure 1: Data structure of NMDP, TMDP, and MDP. (a) In the NMDP, the reward and future observations are determined by all past observation-action pairs. (b) In the TMDP, the reward and future observations depend solely on the current observation-action pairs. (c) In the MDP, the reward and future observations also rely on the current observation-action pairs, with the same-colored slashes indicating identical conditional distributions.

designs consider the classic identically and independently distributed (i.i.d.) setting, failing to account for temporal carryover effects.

Contributions. We make three important contributions below. Firstly, we propose three optimal dynamic treatment allocation strategies in sequential decision making, while accounting for carryover effects over time. The proposed designs are built upon the A-optimality criterion developed in non-dynamic settings (Atkinson et al., 2007), and effectively minimize the variance of the average treatment effect estimator when data follow a (time-invariant) Markov decision process (MDP), a time-varying MDP (TMDP) or a non-Markov decision process (NMDP). Figure 1 offers a visual illustration of these data generating processes. To our knowledge, this represents the first piece of work to provide an analytical form of optimal allocation strategies within these dynamic contexts.

Secondly, by leveraging insights from the established field of off-policy evaluation (OPE, see, for instance, Dudík et al., 2014; Uehara et al., 2022, for reviews), we devise subsequent estimation methodologies that learn the treatment effect estimator using log data produced in accordance with our proposed design. We employ real datasets gathered from a globally recognized ride-sharing company to create a city-scale synthetic environment that mimics the spatio-temporal dynamics of the order dispatching problem. Our findings indicate that the proposed estimators deliver significantly superior accuracy compared to conventional baseline estimators.

Thirdly, we establish the theoretical properties of the proposed designs and estimators. Specifically, we establish that in NMDPs, the optimal design distributes the initial policy randomly, with probabilities corresponding to the standard deviations of the cumulative rewards, and subsequently implements the same policy. However, such a design may not necessarily be optimal in (T)MDPs. To address this, we focus on a category of restricted designs and identify sufficient conditions that ascertain their asymptotic optimalities in (T)MDPs. Moreover, we derive an upper limit for the mean squared error (MSE) of the resulting treatment effect estimator. Extensive simulation studies have been conducted to assess the finite sample performance of the proposed algorithms. Python code implementing the proposed algorithms is available at `https://github.com/tingstat/MDP_design`.

1.1 Related work

Experimental designs. There is an extensive body of literature on experimental design for clinical trials, with a multitude of optimal designs proposed. These include the global treatment balance design (Efron, 1971; Wei, 1977), designs that regulate the marginal covariate distribution across various treatment groups (Pocock and Simon, 1975; Heritier et al., 2005; Liu and Hu, 2022), and criteria-based design anchored on variance minimization (Begg and Iglewicz, 1980; Wong and Zhu, 2008; Chandereng et al., 2020). Other noteworthy designs are the D-optimality and its generalization D_A optimality (Jones and Goos, 2009; Atkinson and Pedrosa, 2017; Loux, 2013; Liu et al., 2022), alongside the A-optimality and its extension A_A optimality (Atkinson et al., 2007; Sverdlov and Rosenberger, 2013; Yin and Zhou, 2017). Moreover, numerous sequential adaptive designs for clinical trials have been developed, including covariate-adaptive designs (Zhu and Hu, 2019; Wang

and Ma, 2021), response-adaptive designs (Zhu et al., 2020; Yu et al., 2022), and covariate-adjusted response-adaptive designs (Zhang et al., 2007; Villar and Rosenberger, 2018). However, these methods were designed for i.i.d. data and thus are not directly applicable to our settings.

Recently, several papers have studied experimental designs with spatial/network spillover effects (Ugander et al., 2013; Baird et al., 2018; Li et al., 2019; Jagadeesan et al., 2020; Nandy et al., 2020; Viviano, 2020; Zhou et al., 2020; Kong et al., 2021; Leung, 2022). Additionally, a few designs have been developed for modern technological companies, such as e-commerce (Bajari et al., 2021; Nandy et al., 2021) and ride-sharing (Johari et al., 2022). However, these studies differ from ours in that they did not utilize NMDP or TMDP models for experimental designs.

Bojinov et al. (2023) explored optimal regular switchback design in the presence of a single experimental unit. Their approach relies on the assumption that the carryover effects last for a fixed duration. However, this assumption is not consistent with our NMDP and (T)MDP models, where the carryover effects, due to state transitions, can potentially last indefinitely.

Finally, Bayesian Optimal Experimental Design (BOED) presents a powerful tool for numerically determining the optimal design (see e.g., Ryan et al., 2016; Rainforth et al., 2023). Recently, works such as Foster et al. (2021), Lim et al. (2022), and Blau et al. (2022) have proposed leveraging deep reinforcement learning to compute the BOED. In contrast, our paper employs a frequentist approach, analytically deriving optimal designs for sequential decision making.

Off-policy evaluation. Our work is closely related to OPE, which seeks to estimate the mean outcome of a new target policy using logged data collected by a different policy. It has been applied in a range of domains including healthcare (Murphy, 2003), education (Mandel et al., 2014), ride-sharing (Shi et al., 2022) and video-sharing (Xu et al., 2023). Recently, it has been employed to conduct A/B testing and mediation analysis in the presence of temporal carryover effects (Hu and Wager, 2022; Farias et al., 2022; Shi et al., 2022; Tang et al., 2022; Ge et al., 2023; Shi et al., 2023). Existing algorithms include value-based methods (Mannor et al., 2007; Antos et al., 2008; Le et al., 2019; Feng et al., 2020; Luckett et al., 2020; Hao et al., 2021; Liao et al., 2021; Chen and Qi, 2022; Shi et al., 2022; Bian et al., 2023), importance (re)sampling-based methods (Swaminathan and Joachims, 2015; Liu et al., 2018; Schlegel et al., 2019; Yin and Wang, 2020; Thams et al., 2021; Wang et al., 2021; Hu and Wager, 2023), and doubly robust or more robust methods (Zhang et al., 2013; Jiang and Li, 2016; Thomas and Brunskill, 2016; Farajtabar et al., 2018; Kallus and Uehara, 2020; Uehara et al., 2020; Shi et al., 2021; Kallus and Uehara, 2022; Liao et al., 2022; Xie et al., 2023). In particular, Kallus and Uehara (2022) and Liao et al. (2022) established the efficiency bounds for OPE, which we leverage to construct optimal designs. However, none of the previous work considered how to design the behavior policy that generates high-quality data to improve the efficiency of the policy value estimator.

Recently, Wan et al. (2022) studied safe experimental designs for efficient policy evaluation under certain safety constraints in a non-dynamic setting without temporal carryover effects. However, it remains unclear how to generalize their methodologies to sequential decision making.

A/B testing. Lastly, our proposal aligns with a body of literature that develops A/B testing algorithms for randomized online experiments (see e.g., Kharitonov et al., 2015; Johari et al., 2017; Yang et al., 2017; Bojinov and Shephard, 2019; Ju et al., 2019; Zhou et al., 2020; Shi et al., 2021; Maharaj et al., 2023; Wan et al., 2023). However, most works focus on the sequential monitoring problem, which involves partitioning the significance level at each interim stage to control the overall type-I error. This is distinct from our goal, which is to generate high-quality experimental data to enhance the efficiency of estimation and inference.

2 Models and Problem Formulation

Data. We consider an experiment spanning n days and each day is divided into T time intervals. On the i-th day ($i = 1, \ldots, n$) and time interval t ($t = 1, \ldots, T$), the decision maker observes certain time-varying market features, denoted as $O_t^{(i)}$ (e.g., the numbers of incoming orders and available drivers in a ride-sharing platform) and chooses to implement one of the two policies. This action is denoted as $A_t^{(i)} \in \{0, 1\}$. By convention, $A_t^{(i)} = 1$ indicates the implementation of a new policy, while $A_t^{(i)} = 0$ signifies the use of the control policy. Afterwards, the decision maker receives an immediate reward $R_t^{(i)}$ and subsequently observes the next observation $O_{t+1}^{(i)}$. This process is

repeated until we reach the termination time T. The observed data can thus be summarized as $\{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}, t = 1, \ldots, T\}_{i=1}^n$.

Model. We consider three models for the aforementioned data generating process: an NMDP, a TMDP, and a (time-invariant) MDP. As depicted in Figure 1, an NMDP is a versatile model that doesn't impose Markovian or stationarity assumptions. The immediate reward (R_t) and future observation (O_{t+1}) depend on the entire history of observations and actions, not merely the current observation-action pair (O_t and A_t). This kind of model has been extensively used in research on learning optimal dynamic treatment regimes (Kosorok and Laber, 2019; Tsiatis et al., 2019). Conversely, a TMDP is a special case of NMDP, with a stronger structural assumption that satisfies Markovanity. It assumes that the current observation-action pair is sufficient to determine the immediate reward and future observation, while allowing this conditional distribution function to vary with time. This type of data generating process is also known as an episodic MDP and has been widely studied in the reinforcement learning literature (see e.g., Jin et al., 2018, 2021; Li et al., 2023). Finally, an MDP is a special TMDP with an additional stationarity assumption.

Objective. A history-dependent policy, denoted as π, is a sequence of decision rules $\{\pi\}_{t\geq 1}$. Each rule, $\pi_t(\bullet|H_t)$, uses the observed data history H_t up to time t as input and provides a probability distribution as output. This output determines the likelihood of choosing each action A_t at time t. The primary objective here is to estimate the Average Treatment Effect (ATE), which is defined as:

$$\text{ATE} = T^{-1}\sum_{t=1}^{T}[\mathbb{E}^1(R_t) - \mathbb{E}^0(R_t)],$$

where $\mathbb{E}^1$ (or $\mathbb{E}^0$) signifies the expected value when the system follows the new (or control) policy. Moreover, for any a in the set $\{0, 1\}$, let's define $V_t^a(h)$ as the value function, $\sum_{k=t}^{T}\mathbb{E}^a(R_k|H_t = h)$, when a certain action is followed. It's evident from this that the ATE equals $T^{-1}\mathbb{E}[V_1^1(O_1) - V_1^0(O_1)]$. Let π^b denote the behavior policy that generates the experimental data, i.e., $\pi_t^b(a|H_t) = \mathbb{P}(A_t = a|H_t)$ for any a and t. Our objective lies in the design of an optimal behavior policy to generate high-quality data so that the mean squared error of the subsequent ATE estimator is minimized.

3 Design, Implementation and Evaluation in NMDPs

We focus on NDMPs in this section. We begin by proposing the optimal dynamic treatment allocation strategy to generate the experimental data. We next discuss some implementation details to implement the proposed design. Finally, we construct the ATE estimator from the data collected.

Design. Jiang and Li (2016) and Kallus and Uehara (2020) established the semiparametric efficiency bound, which is a nonparametric extension of the Cramer-Rao lower bound in parametric models (Bickel et al., 1993), for offline estimation of the average return under a general target policy in NMDPs. In essence, the efficiency bound corresponds to the smallest mean squared error (MSE) among a broad class of regular off-policy estimators. They further developed semiparametrically efficient estimators whose asymptotic MSEs reach this lower bound. Based on their results, one can easily show that the efficiency bound (EB) for the ATE equals:

$$\text{EB}_1(\pi^b) = T^{-2}\sum_{a\in\{0,1\}}\sum_{t=1}^{T}\mathbb{E}^{\pi^b}\Big[\sigma_t(H_t, a)\prod_{k\leq t}\frac{\mathbb{I}(A_k = a)}{\pi_k^b(a|H_k)}\Big]^2 + T^{-2}\text{Var}[V_1^1(O_1) - V_1^0(O_1)], \quad (1)$$

where $\sigma_t^2(H_k, a)$ denotes the conditional variance of the temporal difference error $R_t + V_{t+1}^a(H_{t+1}) - V_t^a(H_t)$ given H_t and that $A_t = a$, and $\mathbb{I}(\bullet)$ denotes the indicator function. Note that the second term on the right-hand-side (RHS) is independent of the behavior policy. However, the first term is a function of π^b in that: (i) the ratio in the first term explicitly involves π^b; (ii) the expectation $\mathbb{E}^{\pi^b}$ is taken with respect to the data generated by π^b.

The proposed design identifies the optimal π^{b*} that minimizes (1). Before delving into the details of the proposed design, we first present the main idea behind it. A key observation is that the cumulative ratio $\prod_{k\leq t}[\mathbb{I}(A_k = a)/\pi_k^b(a|H_k)]$ appears on the RHS of (1) due to the use of sequential importance sampling (see e.g., Jiang and Li, 2016, Equation (4)) to account for the distributional shift between π^b and the global policy that assigns action a at each time. In general, the value of the cumulative ratio grows with the discrepancy between the target and behavior policy. Hence, it is plausible to

Algorithm 1 Treatment allocation algorithm for NMDPs

Input: The burn-in period m_0 for each global policy and the termination day n.

1: Run each global policy for m_0 days and obtain $\{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}\}_{i=1}^{2m_0}$.
2: **while** $2m_0 < m \leq n$ **do**
3: Using the collected data to estimate V_1^a and $\sigma_*(\bullet, a)$ via (3). Obtain $\widehat{\sigma}_*(O_1^{(m)}, a)$.
4: Assign A_1^m according to $\widehat{\pi}_{1,m-1}^{b*}(a|O_1^{(m)}) = \widehat{\sigma}_*(O_1^{(m)}, a)/[\widehat{\sigma}_*(O_1^{(m)}, 1) + \widehat{\sigma}_*(O_1^{(m)}, 0)]$.
5: Set $A_2^{(m)} = \cdots = A_T^{(m)} = A_1^{(m)}$.
6: **end while**

Output: $\{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}\}_{i=1}^{n}$.

expect that an on-policy estimator, where all actions are determined according to the target policy, could potentially minimize the MSE.

We show in Theorem 1 below that the optimal design corresponds to a slight modification of the aforementioned on-policy design. Specifically, it randomly allocates the initial action with probabilities proportional to the standard deviations of the cumulative rewards (see Equation (2)), and assigns the same action afterwards.

Theorem 1. *In NMDPs, π^{b*} that minimizes* (1) *satisfies: (i) for any $a \in \{0, 1\}$*

$$\pi_1^{b*}(a|O_1) = \frac{\sigma_*(O_1, a)}{\sigma_*(O_1, 0) + \sigma_*(O_1, 1)}, \tag{2}$$

where $\sigma_*^2(O_1, a) = \mathbb{E}^a[\{\sum_t (R_t - \mathbb{E}^a R_t)\}^2 | O_1, A_1 = a]$; (ii) $\pi_2^{b*}(A_1|H_2) = \pi_3^{b*}(A_1|H_3) = \cdots = \pi_T^{b*}(A_1|H_T) = 1$ almost surely, or equivalently, $A_1 = A_2 = \cdots = A_T$ under π^{b*}.

Implementation. It remains to estimate the standard deviation σ_* to implement the proposed design. If historical data is available, we can use it to estimate σ_*. Otherwise, we can use data collected from the experiment to adaptively update this parameter. Initially, we run each global policy for m_0 days to generate the data. Next, on the mth day ($2m_0 < m \leq n$), we estimate σ_* using the experimental data collected so far. To this end, we begin by applying existing supervised learning algorithm to estimate the value function V_1^a by regressing the cumulative rewards $\{\sum_t R_t^{(i)} : A_1^{(i)} = a, i < m\}$ on the initial observations $\{O_1^{(i)} : A_1^{(i)} = a, i < m\}$ where the superscript i indicates that the data are collected on the ith day. Let $\widehat{V}_{1,m-1}^a$ denote the resulting estimator. We next employ supervised learning again to regress the squared residuals on the initial observations to estimate $\sigma_*(\bullet, a)$ as

$$\widehat{\sigma}_*(\bullet, a) = \arg\min_{\sigma(\cdot)} \sum_{i=1}^{m-1} \mathbb{I}(A_1^{(i)} = a) \Big\{ \Big[\sum_t R_t^{(i)} - \widehat{V}_{1,m-1}^{(a)}(O_1^{(i)}) \Big]^2 - \sigma^2(O_1^{(i)}) \Big\}^2. \tag{3}$$

Finally, we plug $\widehat{\sigma}_*(O_1^{(m)}, a)$ into the RHS of (2) to obtain $\widehat{\pi}_{1,m-1}^{b*}$, assign $A_1^{(m)}$ according to $\widehat{\pi}_{1,m-1}^{b*}$ and set $A_2^{(m)} = \cdots = A_T^{(m)} = A_1^{(m)}$. We summarize our procedure in Algorithm 1.

Evaluation. Finally, we compute the ATE estimator using the experimental data. Several OPE algorithms, including value-based, importance sampling (IS), and doubly robust (DR) methods, are applicable for this purpose. DR estimators are known for achieving the efficiency bound (Kallus and Uehara, 2020). In our context, we suggest using the following online DR estimator for ATE,

$$\widehat{\mathrm{ATE}}_1 = \sum_{a=0}^{1} \frac{(-1)^{a+1}}{T(n - 2m_0)} \sum_{i=2m_0+1}^{n} \Big[\widehat{V}_{1,i-1}^a(O_1^{(i)}) + \frac{\mathbb{I}(A_1^{(i)} = a)}{\widehat{\pi}_{1,i-1}^{b*}(a|O_1^{(i)})} [\sum_t R_t^{(i)} - \widehat{V}_{1,i-1}^a(O_1^{(i)})] \Big]. \tag{4}$$

Note that both the estimated value function $\widehat{V}_{1,i-1}^a$ and behavior policy $\widehat{\pi}_{1,i-1}^{b*}$ in (4) are computed during the data collection process. These nuisance functions are independent of the data $\{O_1^{(i)}, A_1^{(i)}, R_t^{(i)}\}$ used to construct the policy value. This cross-fitting procedure enables us to circumvent the need to impose certain metric entropy conditions (Díaz, 2020) and has been widely employed in the statistics and machine learning literature (Luedtke and Van Der Laan, 2016; Bibaut et al., 2021; Shi et al., 2021). By design, (4) can be calculated in an *online* manner, eliminating the need to store historical data. Moreover, we implement a burn-in procedure that discards the first $2m_0$

samples to construct $\widehat{\text{ATE}}$. This ensures the consistencies of $\widehat{V}_{1,i}^a$ and $\widehat{\pi}_{1,i}^{b*}$. Concurrently, value-based and IS-based methods are also suitable for estimating ATE. Provided that the nuisance functions are properly modeled, these estimators can achieve the efficiency bound as well (Uehara et al., 2020; Liao et al., 2021; Wang et al., 2021; Shi et al., 2023). The following theorem provides an upper bound for the MSE of the proposed ATE estimator.

Theorem 2. *Suppose that* $\min_i \widehat{\pi}_{1,i}^{b*} \geq \epsilon$ *and* $V_{1,i}^a \leq TR_{\max}$ *for some constants* $\epsilon > 0$ *and* $R_{\max} < \infty$, $\max_{a,i} \mathbb{E}|1/\widehat{\pi}_{1,i}^{b*}(a, O_1) - 1/\pi_1^{b*}(a, O_1)|_2^2 \leq Ci^{-2\alpha_1}$ *and* $\max_{a,i} T^{-2}\mathbb{E}|\widehat{V}_{1,i}^a(O_1) - V_{1,i}^a(O_1)|_2^2 \leq Ci^{-2\alpha_2}$ *for some constants* $C < \infty$, $0 < \alpha_1, \alpha_2 < 1/2$. *Then we have*

$$\mathbb{E}(\widehat{ATE}_1 - ATE)^2 \leq \frac{EB_1(\pi^{b*})}{n - 2m_0} + O(\epsilon^{-1}C(n - 2m_0)^{-1-2\alpha_2}) + O(R_{\max}^2\sqrt{C}(n - 2m_0)^{-1-\alpha_1}).$$

In this case, the exponents α_1 and α_2 denote the convergence rates of the estimated behavior policy and value function, respectively. Note that the first term is the leading term, while the last two terms represent the estimation errors of the nuisance function estimators and converge at a rate much faster than $(n - 2m_0)^{-1}$. The error bound above is minimized when m_0 is zero. In such a scenario, the bound is asymptotically equal to $n^{-1}\text{EB}_1(\pi^{b*})$, which is equivalent to the MSE of an oracle ATE estimator. Specifically, the oracle estimator is a version of the DR estimator with correctly specified value function and behavior policy, and is constructed based on data collected over n days from the optimal design π^{b*}. Thus, the proposed estimator is asymptotically optimal.

Next, we show how to construct the confidence interval of the ATE estimator. A key observation is that, under certain regularity conditions, the proposed estimator is asymptotically normal. More specifically, similar to Kallus and Uehara (2020), we have

$$\sqrt{n - 2m_0}(\widehat{\text{ATE}}_1 - \text{ATE}) \xrightarrow{d} N(0, \text{EB}_1(\pi^{b*})).$$

This motivates us to consider the following Wald-type confidence interval

$$[\widehat{\text{ATE}}_1 - \Phi^{-1}(1 - \alpha/2)\sqrt{\text{EB}_1(\pi^{b*})/(n - 2m_0)}, \widehat{\text{ATE}}_1 + \Phi^{-1}(1 - \alpha/2)\sqrt{\text{EB}_1(\pi^{b*})/(n - 2m_0)}],$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal random variable. It then suffices to estimate the asymptotic variance $\text{EB}_1(\pi^{b*})$ to construct asymptotically valid confidence intervals. Notice that $\widehat{\text{ATE}}_1$ can be represented as an average of martingale differences $\widehat{\text{ATE}}_1 = \sum_{i=2m_0+1}^n \psi_i^1/(n - 2m_0)$ where

$$\psi_i^1 = \sum_{a=0}^{1} \frac{(-1)^{a+1}}{T}\Big[\widehat{V}_{1,i-1}^a(O_1^{(i)}) + \frac{\mathbb{I}(A_1^{(i)} = a)}{\widehat{\pi}_{1,i-1}^{b*}(a|O_1^{(i)})}[\sum_t R_t^{(i)} - \widehat{V}_{1,i-1}^a(O_1^{(i)})]\Big].$$

We propose using the sample variance of $\{\psi_i^1\}_i$ to estimate $\text{EB}_1(\pi^{b*})$. Similar to Theorem 15 of Kallus and Uehara (2022), we can establish the consistency of the sampling variance estimator.

4 Design, Implementation and Evaluation in TMDPs and MDPs

We proceed to study the optimal design and the subsequent ATE estimation in TMDPs.

Design. According to Theorem 2 of Kallus and Uehara (2020), the efficiency bound for the ATE estimator equals

$$\text{EB}_2(\pi^b) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{a\in\{0,1\}} \mathbb{E}^{\pi^b}\Big[\frac{\mathbb{I}(A_t = a)p_t^a(O_t)}{p_t^b(O_t, a)}\sigma_t(O_t, a)\Big]^2 + \frac{1}{T^2}\text{Var}[V_1^1(O_1) - V_1^0(O_1)], \quad (5)$$

where $p_t^1(\bullet)$ ($p_t^0(\bullet)$) denotes the probability mass/density function of O_t under the new policy (control), $p_t^b(\bullet, \bullet)$ denotes the probability mass/density function of the observation-action pair (O_t, A_t) at time t under the behavior policy, and $\sigma_t^2(O_t, a)$ is the conditional variance of the temporal difference error $R_t + V_{t+1}^a(O_{t+1}) - V_t^a(O_t)$ given the current observation O_t and that $A_t = a$.

The efficiency bound in (5) bears a striking resemblance to that in (1). The sole difference lies in the replacement of the sequential IS ratio in (1) with the ratio of the marginal observation-action pair

under the Markov assumption. However, in contrast to NMDPs, the marginal distribution function p_t^b in TMDPs cannot be represented in a closed-form as a function of π^b. This intricate dependence of p_t^b on π^b makes it exceptionally challenging to identify the optimal π^{b*} that minimizes (5). To elaborate, let Π^b represent the class of behavior policies that randomly assign the initial action, and set $\Pi^b = \{\pi^b : \pi_2^b(A_1|H_2) = \pi_3^b(A_1|H_3) = \cdots = \pi_T^b(A_1|H_T) = 1\}$. In Theorem 1, we show that the optimal behavior policy π^{b*} belongs to Π^b in NMDPs. However, this is not the case in TMDPs without additional assumptions.

Proposition 1 (Informal Statement)**.** *There exists a TMDP such that* $\pi^{b*} \notin \Pi^b$.

Identifying the exact optimal design remains challenging. Motivated by the simplicity of the policy class Π^b, we shift our focus to finding the optimal *in-class* behavior policy within Π^b that minimizes (5). We restrict the treatment assignment policies to Π^b for two primary reasons. First, the optimal design for NMDP resides in class Π^b, leveraging this result allows us to anticipate that in-class behavior policies from Π^b will perform well in TMDPs/MDPs, which are subclasses of NMDP; Second, frequently alternating treatments can introduce significant carryover bias in policy evaluation (Hu and Wager, 2022). By focusing on Π^b, we avoid distributional shifts between the behavior policy and the target policy, effectively mitigating carryover bias. To further simplify the analysis, we consider scenarios where the number of decision stages T approaches infinity. The following theorem provides the form of an optimal $\pi^{b*} \in \Pi^b$ that *asymptotically* minimizes (5) under a mild β-mixing condition. Additionally, it shows that π^{b*} is optimal among all history-dependent policies under certain constancy conditions.

Theorem 3. *Suppose the β-mixing condition holds such that* $\lim_{t\to\infty} \mathbb{E}\sup_{a,o} |p_t^a(o|O_1) - p_t^a(o)| \to 0$ *where $p_t^a(\bullet|O_1)$ denotes the probability mass function of O_t given O_1 following the action a. Then an asymptotically optimal in-class behavior policy π^{b*} satisfies (i) for any $a \in \{0, 1\}$,*

$$\pi_1^{b*}(a|O_1) = \pi_1^{b*}(a) = \frac{\sigma_{a*}}{\sigma_{1*} + \sigma_{0*}}, \text{ where } \sigma_{a*}^2 = \sum_{t=1}^{T} \mathbb{E}^a[\sigma_t^2(O_t, 1)]; \tag{6}$$

(ii) $\pi_2^{b*}(A_1|H_2) = \pi_3^{b*}(A_1|H_3) = \cdots = \pi_T^{b*}(A_1|H_T) = 1$ almost surely. In other words, for π^{b*} that satisfies (i) and (ii), we have $\lim_T T[EB_2(\pi^{b*}) - EB_2(\pi^b)] \le 0$ *for any $\pi^b \in \Pi^b$. Additionally, suppose both $\sigma_t^2(o, 1)$ and $\sigma_t^2(o, 0)$ are constant as functions of o and t. Then* $\lim_T T[EB_2(\pi^{b*}) - EB_2(\pi^b)] \le 0$ *for any π^b.*

The β-mixing condition essentially requires the Markov chain to be ergodic. Similar conditions have been imposed in the literature (Bhandari et al., 2018; Zou et al., 2019; Luckett et al., 2020; Kallus and Uehara, 2022; Li et al., 2022; Shi et al., 2022, 2023). It is also weaker than the independence assumption that requires the transition tuples to be independent over time and has been frequently imposed in the literature. The constancy condition is automatically satisfied when the temporal difference error is independent of the current observation and the time step. Different from the optimal design in NMDPs, the initial policy in (6) is independent of the initial observation, thanks to the mixing condition.

Implementation. We summarize the procedure in Algorithm 2.

Evaluation. Similar to Section 3, we can apply value-based, IS-based, or doubly robust estimator to learn the ATE from the collected data. Consider the double reinforcement learning estimator (DRL, Kallus and Uehara, 2020) as an example. A key observation is that the marginal observation-action probability distribution function $p_t^b(O_t, A_t)$ under π^{b*} equals $\pi_1^*(A_t)p_t^{A_t}(O_t)$. As such, the resulting marginal ratio $\mathbb{I}(A_t = a)p_t^a(O_t)/p_t^b(O_t, a)$ is equal to $\mathbb{I}(A_t = a)/\pi_1^*(A_t)$, or equivalently $\mathbb{I}(A_1 = a)/\pi_1^*(A_1)$, independent of O_t. Consequently, one can show the resulting DRL is reduced to

$$\widehat{\text{ATE}}_2 = \sum_{a=0}^{1} \frac{(-1)^{a+1}}{T(n - 2m_0)} \sum_{i=2m_0+1}^{n} \Big\{ \widehat{V}_{1,i-1}^a(O_1^{(i)}) + \frac{\mathbb{I}(A_1^{(i)} = a)}{\widehat{\pi}_{1,i-1}^{b*}(a)} [\sum_t R_t^{(i)} - \widehat{V}_{1,i-1}^a(O_1^{(i)})] \Big\},$$

where $\widehat{V}_{1,i}^a$ and $\widehat{\pi}_1^{b*}(a)$ denotes the estimated value function and estimated initial allocation probability using data from the first ith days; see Step 3 of Algorithm 2 for the estimation procedure. Similar to $\widehat{\text{ATE}}_1$ in (4), $\widehat{\text{ATE}}_2$ can be updated in an online manner without storing historical data. We provide an upper bound for the MSE of $\widehat{\text{ATE}}_2$ below.

Algorithm 2 Treatment allocation algorithm for TMDPs

Input: The burn-in period m_0 for each global policy and the termination day n.

1: Run each global policy for m_0 days and obtain $\{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}\}_{i=1}^{2m_0}$.
2: **while** $2m_0 < m \leq n$ **do**
3: Obtain $\widehat{V}_{t,m-1}^a$ by regressing $\{\sum_{j=t}^T R_t^{(i)} : i < m, A_1^{(i)} = a\}$ on $\{O_t^{(i)} : i < m, A_1^{(a)} = a\}$.
4: For $a \in \{0, 1\}$, set

$$\widehat{\sigma}_{a*}^2 = \frac{\sum_{i<m} \sum_{t=1}^T [R_t^{(i)} + \widehat{V}_{t+1,i}^a(O_{t+1}^{(i)}) - \widehat{V}_{t,i}^a(O_t^{(i)})]^2 \mathbb{I}(A_1^{(i)} = a)}{\sum_{i<m} \mathbb{I}(A_1^{(i)} = a)}.$$

5: Assign $A_1^{(m)}$ according to $\widehat{\pi}_{1,m-1}^{b*}(a) = \widehat{\sigma}_{a*}/(\widehat{\sigma}_{1*} + \widehat{\sigma}_{0*})$.
6: Set $A_2^{(m)} = \cdots = A_T^{(m)} = A_1^{(m)}$.
7: **end while**

Output: $\{O_t^{(i)}, A_t^{(i)}, R_t^{(i)}\}_{i=1}^n$.

Theorem 4. *Suppose that* $\min_i \widehat{\pi}_{1,i}^{b*} \geq \epsilon$ *and* $V_{1,i}^a \leq TR_{\max}$ *for some constants* $\epsilon > 0$ *and* $R_{\max} < \infty$, $\max_{a,i} \mathbb{E}|1/\widehat{\pi}_{1,i}^{b*}(a) - 1/\pi_1^{b*}(a)|_2^2 \leq Ci^{-2\alpha_1}$ *and* $\max_{a,i} T^{-2}\mathbb{E}|\widehat{V}_{1,i}^a(O_1) - V_{1,i}^a(O_1)|_2^2 \leq Ci^{-2\alpha_2}$ *for some constants* $C < \infty$, $0 < \alpha_1, \alpha_2 < 1/2$. *Then we have*

$$\mathbb{E}(\widehat{ATE}_2 - ATE)^2 \leq \frac{EB_2(\pi^{b*})}{n - 2m_0} + O(C\epsilon^{-1}(n - 2m_0)^{-1-2\alpha_2}) + O(\sqrt{C}R_{\max}^2(n - m_0)^{-1-\alpha_1}).$$

Similar to Theorem 2, the MSE of the proposed ATE estimator relies on m_0, the estimated behavior policy and value function. The first term is asymptotically equal to $n^{-1}\text{EB}_2(\pi^{b*})$ when m_0 is much smaller than n. This suggests that the proposed estimator is asymptotically optimal in TDMPs.

Analogous to the procedures in Section 3, we can similarly establish the asymptotic normality of $\widehat{\text{ATE}}_2$, i.e., $\sqrt{n - 2m_0}(\widehat{\text{ATE}}_2 - \text{ATE}) \xrightarrow{d} N(0, \text{EB}_2(\pi^{b*}))$. The corresponding $1 - \alpha$ confidence interval can be constructed by

$$[\widehat{\text{ATE}}_2 - \Phi^{-1}(1 - \alpha/2)\sqrt{\text{EB}_2(\pi^{b*})/(n - 2m_0)}, \widehat{\text{ATE}}_2 + \Phi^{-1}(1 - \alpha/2)\sqrt{\text{EB}_2(\pi^{b*})/(n - 2m_0)}],$$

where the unknown asymptotic variance $\text{EB}_2(\pi^{b*})$ can be similarly estimated via the sampling variance estimator.

MDPs. Finally, we consider MDPs. To save space, we briefly introduce our proposal here and relegate the technical details to the supplementary article. The proposed design is built upon the efficiency bound developed by Liao et al. (2022) in the average reward infinite horizon setting. To simplify the analysis, similar to our proposal in TMDPs, we focus on in-class optimal designs and consider an asymptotic regime where $T \to \infty$. The following theorem summarizes the properties of the proposed behavior policy $\tilde{\pi}^{b*} \in \Pi^b$.

Theorem 5 (Information Statement)**.** *Suppose the* β*-mixing condition holds. Then* $\tilde{\pi}^{b*}$ *asymptotically minimizes the efficiency bound among all* $\pi^b \in \Pi^b$. *In addition, under a constancy condition, it is asymptotically optimal among all policies.*

To learn the ATE, we construct an online doubly robust estimator in the supplementary article based on the estimated relative value functions (Puterman, 2014) and the designed behavior policy.

5 Experiment

This section presents four experiments designed to evaluate different designs. The environments include a tabular example with binary observation variables, an example with continuous observations, a small-scale synthetic dispatch example, and a city-scale real data-based dispatch simulator. We investigate the proposed treatment allocation strategies designed for NMDPs and MDPs. We also implement the following three allocation designs for comparison:

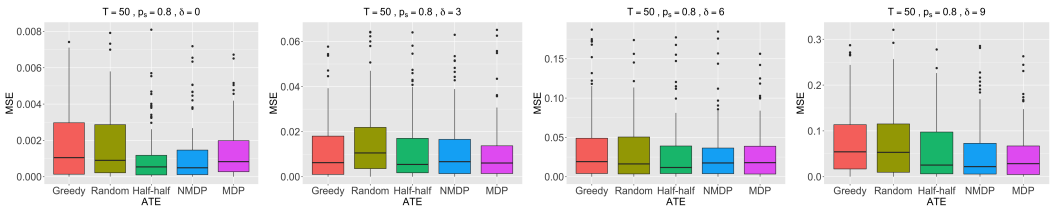

Figure 2: Boxplots of the MSEs of the allocation methods for $p_s = 0.8$ and $T = 50$ in Example 5.1: the four panels correspond to $\delta = 0, 3, 6$, and 9, respectively.

(i) Random: This uniform random treatment allocation design assumes $\mathbb{P}(A_{i,t} = 1) = 1/2$. All $A_{i,t}$s are independent.

(ii) Half-half: This design applies the global treatment for the first $n/2$ days and the global control for the remaining days.

(iii) Greedy: This design uses the ϵ-greedy algorithms that select the current-best treatment, which maximizes the Q-function, with probability $1 - \epsilon$, and employs a uniform random policy with probability ϵ.

We note that (iii) is widely used in the RL literature for online regret minimization. We compare the mean squared errors (MSEs) of the estimates of the average treatment effect based on 200 replicates, and present the results subsequently. We set the burn-in period m_0 to $n/4$ in all the experiments. Detailed information on the data-generating processes can be found in Section S2 of the supplementary material.

Example 5.1 (Binary Observations). In this example, we consider binary observation variables. To better understand the results, we introduce some hyper-parameters that describe the system dynamics. Let p_s denote the marginal probability that the future observation O_{t+1} equals 1 if $A_t = 1$, and it is $1 - p_s$ if $A_t = 0$. Additionally, we use $\delta \in \{0, 3, 6, 9\}$ to characterize the difference of the conditional variance of the reward between the treatment and the control. A larger value of δ indicates a greater difference.

Figure 2 presents boxplots of the MSEs for the allocation methods when $p_s = 0.8$, $T = 50$, and $n = 50$. It is evident that the proposed treatment allocation methods generally outperform the alternatives, offering lower median MSE and smaller variability in most scenarios. However, when $\delta = 0$, the half-half design outperforms ours. This outcome is expected since in this case, the optimal allocation rule assigns the initial action with equal probability, i.e., $\pi_1^{b*} = 0.5$. As δ increases, our proposed estimator surpasses the others in terms of achieving a smaller MSE. Table 1 in the supplementary material presents the means and standard deviations of the MSEs for $T \in \{10, 30, 50\}$ and $p_s \in \{0.5, 0.8\}$. In these scenarios, our proposed designs consistently outperform the alternatives.

Example 5.2 (Continuous Observations). In this example, we use δ_s to denote the difference in the conditional variance of the observation variable between the treatment and the control, and δ to describe the difference of the conditional variance of the reward.

Figure 3 presents the boxplots of all MSEs for $\delta_s = 1$ and $T = 50$ with $n = 50$. Table 2 of the supplementary material includes the Monte Carlo averages and standard errors of MSEs for $T \in \{10, 30, 50\}$ and $\delta_s \in \{0, 1\}$. These results are comparable to those in Example 5.1, demonstrating the superior performance of the proposed design when δ is moderately large. This is expected, as the proposed method takes into account the conditional variance of the temporal difference error.

Example 5.3 (Synthetic Dispatch). Following Xu et al. (2018), we construct a small-scale synthetic dispatch environment to estimate the treatment effect of different order dispatch policies. Specifically, we simulate drivers and orders in a 9×9 spatial grid with a duration of 20 time steps per day. Orders will be canceled if not being responded for a long time. We compare the MDP order dispatch strategy (Xu et al., 2018) against the distance-based dispatch method which minimizes the total distance between drivers and passengers. The reward of interest is given by the total revenue. The observation variables are set to the number of orders and the number of drivers at each time. The number of days is set to be $n \in \{30, 50, 100\}$. All the methods are tested using 100 orders with the number of drivers being either generated from the uniform distribution $U(25, 30)$, or being fixed to 25, 50. A detailed description of the environment can be found in the supplementary document.

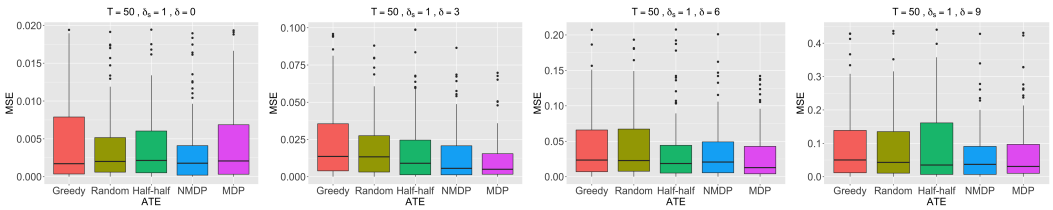

Figure 3: Boxplots of the MSEs of the allocation methods corresponding to $\delta_s = 1$ and $T = 50$ in Example 5.2: the four panels correspond to $\delta = 0, 3, 6, 9$, respectively.

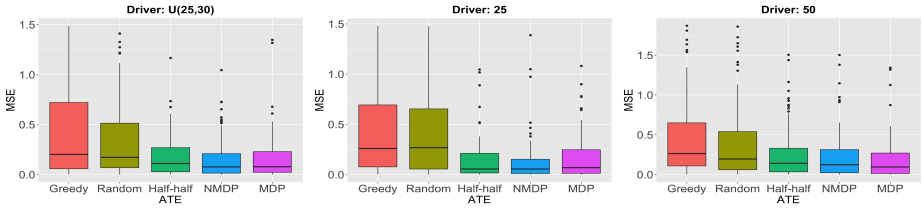

Figure 4: Boxplots of the MSEs of the allocation methods in Example 5.3 with $n = 50$: the three panels correspond to drivers generated from $U(25, 30)$ or fixed as 25 and 50, respectively.

Figure 4 presents the boxplots of MSEs for $n = 50$ with different numbers of drivers. Table 3 of the supplementary material reports the values of MSEs of the four methods. As expected, the MSEs of all the methods decrease as the increase of the number of days n. The proposed method achieves the best performance in almost all scenarios.

Example 5.4 (Real-Data Based Dispatch). We evaluate the proposed treatment allocation method on a dispatch simulator based on a city-scale order-driver historical dataset from a world-leading ride-sharing platform. The dataset consists of temporal spatial information of orders or drivers and numerical features of them. We adopt the simulator in Tang et al. (2019) to generate data based on the historical dataset. The distributions of drivers and orders are set to be identical to the distributions of historical data. Similar to Example 5.3, the two policies being compared are the MDP strategy and the distance-based dispatch method. The reward of interest is the total Gross Merchandise Volume (GMV), and the observation variables include the number of drivers and the number of orders.

Since the ground truth of the average treatment effect is large, we calculate the relative MSE of the estimated average treatment effect for each method. RMSEs of ATE estimates for distinct allocation designs are reported in Figure 5 with the number of days $n = 7$ and $n = 10$. It can be seen that the proposed method outperforms all its counterparts. Meanwhile, as the episode n gets larger, the RMSEs of most allocation designs become smaller.

Acknowledgement

Li's research is partially supported by the National Science Foundation of China 12101388, CCF-DiDi GAIA Collaborative Research Funds for Young Scholars and Program for Innovative Research Team of Shanghai University of Finance and Economics. Shi's research is partially supported by an EPSRC grant EP/W014971/1. Zhou's work is partially supported by National Natural Science Foundation of China 12001356, "Chenguang Program" supported by Shanghai Education Development

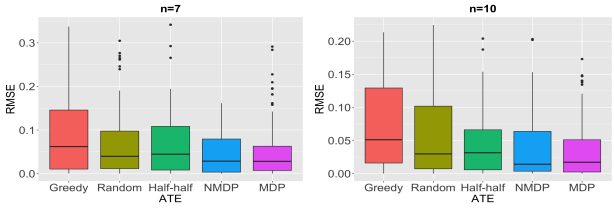

Figure 5: Boxplots of the RMSEs of the allocation methods in Example 5.4 with $n = 7$ and $n = 10$.

Foundation and Shanghai Municipal Education Commission, Open Research Projects of Zhejiang Lab NO.2022RC0AB06, Shanghai Research Center for Data Science and Decision Technology. We thank the anonymous referees and the meta reviewer for their constructive comments, which have led to a significant improvement of the earlier version of this article.

References

Antos, A., C. Szepesvári, and R. Munos (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning 71*, 89–129.

Atkinson, A., A. Donev, and R. Tobias (2007). *Optimum experimental designs, with SAS*, Volume 34. OUP Oxford.

Atkinson, A. and D. Pedrosa (2017). Optimum design and sequential treatment allocation in an experiment in deep brain stimulation with sets of treatment combinations. *Statistics in Medicine 36*(30), 4804–4815.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics 100*(5), 844–860.

Bajari, P., B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. Richardson, and I. M. Rosen (2021). Multiple randomization designs. *arXiv preprint arXiv:2112.13495*.

Begg, C. B. and B. Iglewicz (1980). A treatment allocation procedure for sequential clinical trials. *Biometrics 36*(1), 81–90.

Bhandari, J., D. Russo, and R. Singal (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pp. 1691–1692. PMLR.

Bian, Z., C. Shi, Z. Qi, and L. Wang (2023). Off-policy evaluation in doubly inhomogeneous environments. *arXiv preprint arXiv:2306.08719*.

Bibaut, A., M. Dimakopoulou, N. Kallus, A. Chambaz, and M. van Der Laan (2021). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems 34*, 28548–28559.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Springer.

Blau, T., E. V. Bonilla, I. Chades, and A. Dezfouli (2022). Optimizing sequential experimental design with deep reinforcement learning. In *International Conference on Machine Learning*, pp. 2107–2128. PMLR.

Bojinov, I. and N. Shephard (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association 114*(528), 1665–1682.

Bojinov, I., D. Simchi-Levi, and J. Zhao (2023). Design and analysis of switchback experiments. *Management Science 69*(7), 3759–3777.

Chandereng, T., X. Wei, and R. Chappell (2020). Imbalanced randomization in clinical trials. *Statistics in Medicine 39*(16), 2185–2196.

Chen, X. and Z. Qi (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pp. 3558–3582. PMLR.

Díaz, I. (2020). Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics 21*(2), 353–358.

Dudík, M., D. Erhan, J. Langford, and L. Li (2014). Doubly robust policy evaluation and optimization. *Statistical Science 29*(4), 485–511.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika 58*(3), 403–417.

Farajtabar, M., Y. Chow, and M. Ghavamzadeh (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR.

Farias, V., A. Li, T. Peng, and A. Zheng (2022). Markovian interference in experiments. *Advances in Neural Information Processing Systems 35*, 535–549.

Feng, Y., T. Ren, Z. Tang, and Q. Liu (2020). Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR.

Foster, A., D. R. Ivanova, I. Malik, and T. Rainforth (2021). Deep adaptive design: Amortizing sequential bayesian experimental design. In *International Conference on Machine Learning*, pp. 3384–3395. PMLR.

Ge, L., J. Wang, C. Shi, Z. Wu, and R. Song (2023). A reinforcement learning framework for dynamic mediation analysis. In *International Conference on Machine Learning*. PMLR.

Hao, B., X. Ji, Y. Duan, H. Lu, C. Szepesvári, and M. Wang (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR.

Heritier, S., V. Gebski, and A. Pillai (2005). Dynamic balancing randomization in controlled clinical trials. *Statistics in Medicine 24*(24), 3729–3741.

Hu, Y. and S. Wager (2022). Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*.

Hu, Y. and S. Wager (2023). Off-policy evaluation in partially observed markov decision processes under sequential ignorability.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR.

Jin, C., Z. Allen-Zhu, S. Bubeck, and M. I. Jordan (2018). Is Q-learning provably efficient? *Advances in Neural Information Processing Systems 31*.

Jin, Y., Z. Yang, and Z. Wang (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (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.

Johari, R., H. Li, I. Liskovich, and G. Y. Weintraub (2022). Experimental design in two-sided platforms: An analysis of bias. *Management Science 68*(10), 7065–7791.

Jones, B. and P. Goos (2009). D-optimal design of split-split-plot experiments. *Biometrika 96*(1), 67–82.

Ju, N., D. Hu, A. Henderson, and L. Hong (2019). A sequential test for selecting the better variant: Online A/B testing, adaptive allocation, and continuous monitoring. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 492–500.

Kallus, N. and M. Uehara (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research 21*(1), 6742–6804.

Kallus, N. and M. Uehara (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research 70*(6), 3282–3302.

Kharitonov, E., A. Vorobev, C. Macdonald, P. Serdyukov, and I. Ounis (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.

Kong, X., M. Yuan, and W. Zheng (2021). Approximate and exact designs for total effects. *The Annals of Statistics 49*(3), 1594–1625.

Kosorok, M. R. and E. B. Laber (2019). Precision medicine. *Annual Review of Statistics and its Application 6*, 263–286.

Le, H., C. Voloshin, and Y. Yue (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR.

Leung, M. P. (2022). Rate-optimal cluster-randomized designs for spatial interference. *The Annals of Statistics 50*(5), 3064–3087.

Li, G., Y. Yan, Y. Chen, and J. Fan (2023). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*.

Li, M., C. Shi, Z. Wu, and P. Fryzlewicz (2022). Testing stationarity and change point detection in reinforcement learning. *arXiv preprint arXiv:2203.01707*.

Li, X., P. Ding, Q. Lin, D. Yang, and J. S. Liu (2019). Randomization inference for peer effects. *Journal of the American Statistical Association 114*(528), 1651–1664.

Liao, P., P. Klasnja, and S. Murphy (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association 116*(533), 382–391.

Liao, P., Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy (2022). Batch policy learning in average reward markov decision processes. *The Annals of Statistics 50*(6), 3364–3387.

Lim, V., E. Novoseller, J. Ichnowski, H. Huang, and K. Goldberg (2022). Policy-based bayesian experimental design for non-differentiable implicit models. *arXiv preprint arXiv:2203.04272*.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems 31*.

Liu, X., R.-X. Yue, and W. Kee Wong (2022). Equivalence theorems for c and da-optimality for linear mixed effects models with applications to multitreatment group assignments in health care. *Scandinavian Journal of Statistics 49*(4), 1842–1859.

Liu, Y. and F. Hu (2022). Balancing unobserved covariates with covariate-adaptive randomized experiments. *Journal of the American Statistical Association 117*(538), 875–886.

Loux, T. M. (2013). A simple, flexible, and effective covariate-adaptive treatment allocation procedure. *Statistics in Medicine 32*(22), 3775–3787.

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association 115*(530), 692–706.

Luedtke, A. R. and M. J. Van Der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics 44*(2), 713.

Maharaj, A., R. Sinha, D. Arbour, I. Waudby-Smith, S. Z. Liu, M. Sinha, R. Addanki, A. Ramdas, M. Garg, and V. Swaminathan (2023). Anytime-valid confidence sequences in an enterprise A/B testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 396–400.

Mandel, T., Y.-E. Liu, S. Levine, E. Brunskill, and Z. Popovic (2014). Offline policy evaluation across representations with applications to educational games. In *AAMAS*, Volume 1077.

Mannor, S., D. Simester, P. Sun, and J. N. Tsitsiklis (2007). Bias and variance approximation in value function estimates. *Management Science 53*(2), 308–322.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 331–355.

Nandy, P., K. Basu, S. Chatterjee, and Y. Tu (2020). A/B testing in dense large-scale networks: design and inference. *Advances in Neural Information Processing Systems 33*, 2870–2880.

Nandy, P., D. Venugopalan, C. Lo, and S. Chatterjee (2021). A/B testing for recommender systems in a two-sided marketplace. *Advances in Neural Information Processing Systems 34*, 6466–6477.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics 31*(1), 103–115.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rainforth, T., A. Foster, D. R. Ivanova, and F. B. Smith (2023). Modern bayesian experimental design. *arXiv preprint arXiv:2302.14545*.

Ryan, E. G., C. C. Drovandi, J. M. McGree, and A. N. Pettitt (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review 84*(1), 128–154.

Schlegel, M., W. Chung, D. Graves, J. Qian, and M. White (2019). Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems 32*.

Shi, C., S. Luo, H. Zhu, and R. Song (2021). An online sequential test for qualitative treatment effects. *The Journal of Machine Learning Research 22*(1), 13061–13111.

Shi, C., Z. Qi, J. Wang, and F. Zhou (2023). Value enhancement of reinforcement learning via efficient and robust trust region optimization. *Journal of the American Statistical Association accepted*.

Shi, C., R. Song, W. Lu, and R. Li (2021). Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association 116*(535), 1307–1318.

Shi, C., R. Wan, V. Chernozhukov, and R. Song (2021). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pp. 9580–9591. PMLR.

Shi, C., R. Wan, G. Song, S. Luo, H. Zhu, and R. Song (2022). A multi-agent reinforcement learning framework for off-policy evaluation in two-sided markets. *Annals of Applied Statistics accepted*.

Shi, C., X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song (2023). Dynamic causal effects evaluation in A/B testing with a reinforcement learning framework. *Journal of the American Statistical Association 118*(543), 2059–2071.

Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 765–793.

Shi, C., J. Zhu, S. Ye, S. Luo, H. Zhu, and R. Song (2022). Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association (accepted)*.

Sverdlov, O. and W. F. Rosenberger (2013). On recent advances in optimal allocation designs in clinical trials. *Journal of Statistical Theory and Practice 7*, 753–773.

Sverdlov, O., Y. Ryeznik, and W. K. Wong (2020). On optimal designs for clinical trials: an updated review. *Journal of Statistical Theory and Practice 14*(1), 10.

Swaminathan, A. and T. Joachims (2015). The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems 28*.

Tang, X., Z. Qin, F. Zhang, Z. Wang, Z. Xu, Y. Ma, H. Zhu, and J. Ye (2019). A deep value-network based approach for multi-driver order dispatching. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1780–1790.

Tang, Z., Y. Duan, S. Zhang, and L. Li (2022). A reinforcement learning approach to estimating long-term treatment effects. *arXiv preprint arXiv:2210.07536*.

Thams, N., S. Saengkyongam, N. Pfister, and J. Peters (2021). Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*.

Thomas, P. and E. Brunskill (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR.

Tsiatis, A. A., M. Davidian, S. T. Holloway, and E. B. Laber (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC press.

Uehara, M., J. Huang, and N. Jiang (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR.

Uehara, M., C. Shi, and N. Kallus (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*.

Ugander, J., B. Karrer, L. Backstrom, and J. Kleinberg (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 329–337.

Villar, S. S. and W. F. Rosenberger (2018). Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking gittins index rule. *Biometrics 74*(1), 49–57.

Viviano, D. (2020). Experimental design under network interference. *arXiv preprint arXiv:2003.08421*.

Wan, R., B. Kveton, and R. Song (2022). Safe exploration for efficient policy evaluation and comparison. In *International Conference on Machine Learning*, pp. 22491–22511. PMLR.

Wan, R., Y. Liu, J. McQueen, D. Hains, and R. Song (2023). Experimentation platforms meet reinforcement learning: Bayesian sequential decision-making for continuous monitoring. *arXiv preprint arXiv:2304.00420*.

Wang, J., Z. Qi, and R. K. Wong (2021). Projected state-action balancing weights for offline reinforcement learning. *arXiv preprint arXiv:2109.04640*.

Wang, T. and W. Ma (2021). The impact of misclassification on covariate-adaptive randomized clinical trials. *Biometrics 77*(2), 451–464.

Wei, L.-J. (1977). A class of designs for sequential clinical trials. *Journal of the American Statistical Association 72*(358), 382–386.

Wong, W. K. and W. Zhu (2008). Optimum treatment allocation rules under a variance heterogeneity model. *Statistics in Medicine 27*(22), 4581–4595.

Xie, C., W. Yang, and Z. Zhang (2023). Semiparametrically efficient off-policy evaluation in linear markov decision processes. In *International Conference on Machine Learning*. PMLR.

Xu, Y., J. Zhu, C. Shi, S. Luo, and R. Song (2023). An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pp. 38848–38880. PMLR.

Xu, Z., Z. Li, Q. Guan, D. Zhang, Q. Li, J. Nan, C. Liu, W. Bian, and J. Ye (2018). Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 905–913.

Yang, F., A. Ramdas, K. G. Jamieson, and M. J. Wainwright (2017). A framework for multi-A(rmed)/B(andit) testing with online FDR control. *Advances in Neural Information Processing Systems 30*.

Yin, M. and Y.-X. Wang (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR.

Yin, Y. and J. Zhou (2017). Optimal designs for regression models using the second-order least squares estimator. *Statistica Sinica 27*(4), 1841–1856.

Yu, Y., C. Xu, J. Zhong, and S. H. Cheung (2022). Comparison of treatments with ordinal responses in trials with sequential monitoring and response-adaptive randomization. *Statistics in Medicine 41*(25), 5061–5083.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika 100*(3), 681–694.

Zhang, L.-X., F. Hu, S. H. Cheung, and W. S. Chan (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics 35*(3), 1166–1182.

Zhou, Y., Y. Liu, P. Li, and F. Hu (2020). Cluster-adaptive network A/B testing: From randomization to estimation. *arXiv preprint arXiv:2008.08648*.

Zhu, H. and F. Hu (2019). Sequential monitoring of covariate-adaptive randomized clinical trials. *Statistica Sinica 29*(1), 265–282.

Zhu, H., J. Piao, J. J. Lee, F. Hu, and L. Zhang (2020). Response adaptive randomization procedures in seamless phase ii/iii clinical trials. *Journal of Biopharmaceutical Statistics 30*(1), 3–17.

Zou, S., T. Xu, and Y. Liang (2019). Finite-sample analysis for sarsa with linear function approximation. *Advances in Neural Information Processing Systems 32*.