Spatiotemporal Causal Effects Evaluation: A Multi-Agent Reinforcement Learning Framework

Anonymous Author(s)
Affiliation
Address
email

Abstract

Online experiment is a default option for technological companies to make data-driven product decisions. Major challenge arise in experiments where multiple units in different areas receive sequences of treatments over time. Causal effects evaluation is extremely challenging in those experiments because (i) spatial and temporal proximities induce interference between locations and times; (ii) the large number of locations results in the curse of dimensionality; (iii) the short duration of the experiment leads to data scarcity. In this paper, we introduce a multi-agent reinforcement learning framework for carrying spatiotemporal causal effects evaluation and propose novel estimators for mean outcomes under different products that are consistent despite the high-dimensionality of state-action space. The proposed estimator works favourably in simulation experiments. We further illustrate our method using data from a ridesharing company to evaluate the effects of applying subsidizing policies in different areas.

1 Introduction

Online experiment is a standard business strategy for technological companies to make data-driven product decisions. The existing literature on causal inference has mostly focused on the setting where no interference occurs, i.e., the outcome of each experimental unit depends only on its treatment status. This assumption is referred to as the stable unit treatment value assumption [SUTVA, 21, 22].

In many experiments, however, there are multiple units in different areas that receive sequences of treatments over time. For instance, suppose a ride-sharing company would like to evaluate the effect of applying different subsidizing policies to drivers in different spatial units of a city. Implementing a subsidizing policy at one location will attract drivers from other areas to that location, thus affecting the spatial distribution of drivers in the city. Consequently, the subsidizing policy at one location will impact outcomes of other areas, inducing interference between spatial units. In addition, the subsidizing policy at a given time will affect both current and future rewards, inducing interference over time. This leads to the violation of SUTVA.

Contribution. The focus of this paper is to evaluate the impact of multiple products in the presence of spatiotemporal interference, using data from online experiment. Our contributions are multi-fold. First, we introduce a multi-agent reinforcement learning [MARL, see e.g., 19] framework for spatiotemporal causal effects evaluation. Each spatial unit in the city is considered as an agent. In addition to the treatment-outcome pairs, it is assumed that each agent is associated with a set of time-varying confounding variables. This naturally leads to a multi-agent system. Under this framework, the carryover effects in space is modeled by the interactions between different agents. The carryover effects in time is modeled by the dynamic system transitions. See the causal diagram depicted in Figure 1 for an illustration. Estimation of the mean outcome under different products is reduced to the off-policy evaluation problem in MARL. This addresses the challenge on spatiotemporal interference.

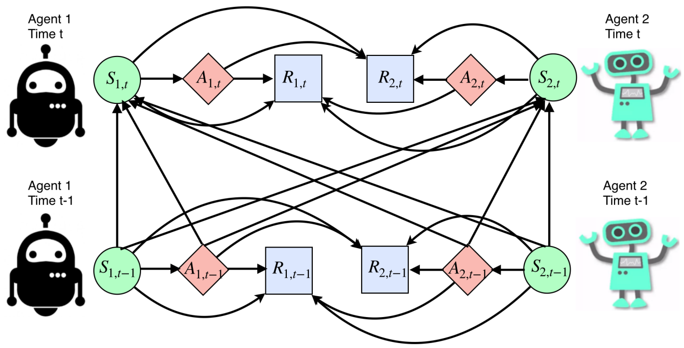

Figure 1: Causal diagram for a multi-agent system with two agents. $(S_{j,t}, A_{j,t}, R_{j,t})$ represents the state-treatment-outcome triplet of the j-th agent at time t.

Second, we propose a novel off-policy evaluation procedure in MARL. The proposed estimator requires estimation of the density ratio of the stationary state distribution and the Q-function associated with each single agent. The key ingredient of our method lies in learning the density ratio and Q-function based on mean-field approximation (see Section 3 for details) and aggregating these estimators properly to satisfy the doubly-robustness property. The mean field approximation effectively reduces the high-dimensional state-action space to a moderate scale, leading to a value estimator with decreased variance. The doubly-robustness guarantees our estimated value is consistent when either the density ratio or the Q-function is well-approximated, reducing its bias resulting from the mean-field approximation. This addresses the challenge on the curse of dimensionality.

Third, we rigorously investigate the statistical properties of our estimator. In particular, we establish its doubly-robustness property (Theorem 2) and derive its "oracle" property when both the density ratio and the Q-function are well approximated (Theorem 3). To prove these results, we develop an exponential inequality for suprema of empirical processes under weak dependence (Lemma B.1), which is useful for finite-sample analysis of machine learning estimates based on dependent observations. Our theory allows the number of spatial units N to be either bounded or diverge to infinity. As such, the proposed estimator offers a useful policy evaluation tool to a wide range of applications in the presence of spatiotemporal interference.

Related work. There is a huge literature on causal inference. As commented before, most works considered settings without interference. Our work is related to research on space- or time-dependent casual effects evaluation [see e.g., 10, 24, 26, 7, 1, 5, 3, 4, 18]. However, none of the above cited works studied the interference effects in both space and time. In addition, the reinforcement learning framework has not been utilized in these papers to characterize the casual effects.

In addition to the literature on causal inference, our work is also related to a line of research on MARL in the cooperative setting [see e.g., 30, for an overview] and online control of infectious diseases [see e.g., 13]. Most works in the literature considered the *policy optimization* problem where the objective is that agents collaborate to optimize a long-term reward. In particular, [28] developed a mean field Q-learning algorithms in the discounted-reward setting. We remark that *policy evaluation* is an ultimately different problem as policy optimization. On the other hand, the proposed value estimator relies on an estimated Q-function. To this end, we extend [28]'s proposal to the average-reward setting, which is more suitable for our application. In addition, we propose a mean field algorithm to approximate the density ratio of the stationary state distribution in a multi-agent system, in order to construct our doubly-robust estimates.

Furthermore, our work is also related to the literature on off-policy evaluation in reinforcement learning, in particular, on importance-sampling based or doubly-robust estimation of the value [see e.g., 25, 11, 15, 12, 27, 23]. However, all the above cited papers considered a single-agent system, which is ultimately different from our setup.

2 Causality and MARL

In this section, we extend Robin's potential outcome framework to the multi-agent system. We assume there are only two treatments (actions) associated with the i-th spatial unit (agent), i.e. the action space $\mathcal{A}_i = \{0, 1\}$. In our ride-sharing applications, the two treatments correspond to applying the subsidizing policy to a given area or not. For $1 \le i \le N$, let $\mathbb{S}_i$ denote the state space associated with the i-th agent. In addition, let $\mathbb{S}_0$ denote the space of some global state variables in the system (such as time of day in our applications). The joint state and action spaces are given by $\mathbb{S} = \mathbb{S}_0 \times \mathbb{S}_1 \times \mathbb{S}_2 \times \cdots \times \mathbb{S}_N$ and $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N = \{0,1\}^N$, respectively.

For a sequence of N-dimensional vectors $\boldsymbol{a}_0, \boldsymbol{a}_1, \cdots, \boldsymbol{a}_t \in \{0,1\}^N$, define a treatment history vector $\bar{\boldsymbol{a}}_t = (\boldsymbol{a}_0^\top, \boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_t^\top)^\top$ up to time t. For each $i \in \{1, \cdots, N\}$, let $S_{i,t}^*(\bar{\boldsymbol{a}}_{t-1}) \in \mathcal{S}_i$ and $R_{i,t}^*(\bar{\boldsymbol{a}}_t) \in \mathbb{R}$ be the potential state and reward (outcome) associated with the i-th agent at time t, that would occur had all agents followed $\bar{\boldsymbol{a}}_t$. Similarly, let $S_{0,t}^*(\bar{\boldsymbol{a}}_{t-1})$ be the potential global state that would occur at time t had all agents followed $\bar{\boldsymbol{a}}_{t-1}$. Note that different action histories would lead to different potential outcomes. More importantly, these potential outcomes cannot be directly observed. We only have access to those following actions selected by the agents (see Condition (CA) below). Consequently, causal inference is inherently a missing data problem. More specifically, let $\{(S_{0,t}, S_{i,t}, A_{i,t}, R_{i,t}) : 1 \le i \le N, 0 \le t < T\} \cup \{S_{i,T} : 0 \le i \le N\}$ be the observed data where $S_{0,t}$ denotes the observed global state at time t, $(S_{i,t}, A_{i,t}, R_{i,t})$ stands for the observed state-action-reward triplet associated with the i-th agent at time t and T is the termination time of the experiment. Let $\boldsymbol{A}_t = (A_{1,t}, \cdots, A_{N,t})^\top$ be the observed treatments at time t and $\bar{\boldsymbol{A}}_t = (\boldsymbol{A}_0^\top, \boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_t^\top)^\top$. We make the following consistency assumption (CA).

(CA) $S_{i,t} = S_{i,t}^*(\bar{\boldsymbol{A}}_{t-1})$, $R_{i,t} = R_{i,t}^*(\bar{\boldsymbol{A}}_t)$ almost surely for any i and t.

SUTVA requires $S_{i,t+1}^*$ and $R_{i,t}^*$ to be functions of $A_{i,t}$ only. The above condition extends SUTVA to settings with spatiotemporal interference. Specifically, these potential outcomes are allowed to depend on not only past treatments, but actions selected by other agents as well. The following sequential randomization assumption (SRA) guarantees our causal estimands are identifiable from the observed data.

(SRA) $\boldsymbol{A}_t \perp\!\!\!\perp \boldsymbol{W}^* | \{(S_{0,j}, S_{i,j}, A_{i,j}, R_{i,j}) : 1 \le i \le N, 0 \le j < t\} \cup \{S_{i,t} : 0 \le i \le N\}$ for any t where $\boldsymbol{W}^* = \cup_{t \ge 0, \bar{\boldsymbol{a}}_t \in \{0,1\}^{N(t+1)}} \boldsymbol{W}_t^*(\bar{\boldsymbol{a}}_t)$ where $\boldsymbol{W}_t^*(\bar{\boldsymbol{a}}_t)$ denotes the set of potential outcomes following $\bar{\boldsymbol{a}}_t$ up to time t, i.e.,

$$\boldsymbol{W}_t^*(\bar{\boldsymbol{a}}_t) = \{(S_{0,j}^*(\bar{\boldsymbol{a}}_{j-1}), S_{i,j}^*(\bar{\boldsymbol{a}}_{j-1}), R_{i,j}^*(\bar{\boldsymbol{a}}_j)) : 1 \le i \le N, 0 \le j \le t\}.$$

SRA is satisfied in our applications where $A_{i,t}$'s are i.i.d. Bernoulli random variables, independent of other observations. More generally, it automatically holds in randomized experiments where the distribution of $\boldsymbol{A}_t$ is completely determined by the observed state-action-reward history. However, this condition cannot be verified from data from observational studies. We note that CA and SRA are commonly imposed in sequential decision making problems [see e.g., 17, 20, 29, 8, 13, 16].

Let $\boldsymbol{S}_t^*(\bar{\boldsymbol{a}}_{t-1}) = \{S_{0,t}^*(\bar{\boldsymbol{a}}_{t-1}), S_{1,t}^*(\bar{\boldsymbol{a}}_{t-1}), \cdots, S_{N,t}^*(\bar{\boldsymbol{a}}_{t-1})\}^\top$ be the potential state vector at time t. Next we introduce the Markov assumption (MA) and the conditional mean independence assumption (CMIA). These conditions assume the system dynamics are homogeneous over time, enabling consistent estimation of our causal estimands.

(MA) There exists a Markov transition kernel $\mathcal{P} : \mathbb{S} \times \mathcal{A} \times \mathbb{S} \to \mathbb{R}$ such that for any $t \ge 0$, $\bar{\boldsymbol{a}}_t \in \{0,1\}^{N(t+1)}$ and $\mathcal{S} \in \mathbb{S}$, we have

$$\mathbb{P}\{\boldsymbol{S}_t^*(\bar{\boldsymbol{a}}_{t-1}) \in \mathcal{S} | \boldsymbol{W}_{t-1}^*(\bar{\boldsymbol{a}}_{t-1})\} = \mathcal{P}(\mathcal{S}; \boldsymbol{a}_{t-1}, \boldsymbol{S}_{t-1}^*(\bar{\boldsymbol{a}}_{t-2})).$$

(CMIA) There exist functions $r_1, \cdots, r_N$ such that for any $1 \le i \le N$, $t \ge 0$, $\bar{\boldsymbol{a}}_t \in \{0,1\}^{N(t+1)}$,

$$\mathbb{E}\{R_{i,t}^*(\bar{\boldsymbol{a}}_t) | \boldsymbol{S}_t^*(\bar{\boldsymbol{a}}_{t-1}), \boldsymbol{W}_{t-1}^*(\bar{\boldsymbol{a}}_{t-1})\} = r_i(\boldsymbol{a}_t, \boldsymbol{S}_t^*(\bar{\boldsymbol{a}}_{t-1})).$$

Throughout this paper, we assume CA, SRA, MA and CMIA hold.

We next describe our causal estimands. We focus on the class of non-dynamic policies indexed by an N-dimensional vector $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)^\top \in \{0,1\}^N$. Under $\boldsymbol{\pi}$, the i-th spatial unit will receive the same treatment π_i over time. We are interested in evaluating the average reward under $\boldsymbol{\pi}$. In our applications, this helps the company to decide whether to apply subsidizing policies to specific areas in a given city according to $\boldsymbol{\pi}$ or not, under some budget constraints. To this end, we present the average treatment effect (ATE) for multi-agent systems in the following definition.

Definition (ATE). Given a control policy $\boldsymbol{\pi}_0$ and a new policy $\boldsymbol{\pi}_1$, ATE is defined as the difference between their long term values,

$$\text{ATE}(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1) = \lim_{t \to \infty} \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t} \mathbb{E} R_{i,j}^*(\boldsymbol{\pi}_1) - \lim_{t \to \infty} \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t} \mathbb{E} R_{i,j}^*(\boldsymbol{\pi}_0), \tag{1}$$

where $\boldsymbol{S}_t^*(\boldsymbol{\pi})$ and $R_{i,t}^*(\boldsymbol{\pi})$ denote the potential outcomes that would occur at time t had all agents followed the non-dynamic policy $\boldsymbol{\pi}$.

We focus on comparing a standard policy $\boldsymbol{\pi}_0$ with several new alternatives $\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_m\}$. This requires to evaluate ATE$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_\ell)$ for $\ell = 1, 2, \cdots, m$. By definition, it is equivalent to evaluate the value under each $\boldsymbol{\pi}_\ell$, i.e., $V(\boldsymbol{\pi}_\ell) = \lim_{t\to\infty}(Nt)^{-1}\sum_{i=1}^N\sum_{j=0}^t \mathbb{E}R_{i,j}^*(\boldsymbol{\pi}_\ell)$. Note that $V(\boldsymbol{\pi}_\ell)$ can be represented by $N^{-1}\sum_{i=1}^N V_i(\boldsymbol{\pi}_\ell)$ where $V_i(\boldsymbol{\pi}_\ell) = \lim_{t\to\infty} t^{-1}\sum_{j=0}^t \mathbb{E}R_{i,j}^*(\boldsymbol{\pi}_\ell)$. It suffices to estimate $V_i(\boldsymbol{\pi}_\ell)$ for $i \in \{1, 2, \cdots, N\}$. We detail our procedure in the next section.

3 Off-policy evaluation in MARL

To better illustrate the idea, we begin by proposing an importance-sampling (IS) based estimator for $V_i(\boldsymbol{\pi})$. A doubly-robust version is presented later in this section. Let $\boldsymbol{S}_t = (S_{0,t}^\top, S_{1,t}^\top, \cdots, S_{N,t}^\top)^\top$. We first introduce some assumptions.

(A1) The system follows a stationary behavior policy b, i.e., $\mathbb{P}(\boldsymbol{A}_t = \boldsymbol{a}_t | \{S_{0,j}, A_{i,j}, S_{i,j}, R_{i,j}\}_{1\le i\le N, 0\le j<t} \cup \{S_{i,t}\}_{0\le i\le N}) = b(\boldsymbol{a}_t|\boldsymbol{S}_t), \quad \forall \boldsymbol{a}_t \in \{0,1\}^N$.

(A2) The process $\{(\boldsymbol{S}_t, \boldsymbol{A}_t) : t \ge 0\}$ is strictly stationary. Its β-mixing coefficients $\{\beta(q) : q \ge 0\}$ [see e.g., 6, for a detailed definition] satisfy $\beta(q) \le \kappa_0\rho^q$ for some constants $\kappa_0 > 0$, $0 < \rho < 1$.

(A1) implies that $\boldsymbol{A}_t$ depends on past observations only through $\boldsymbol{S}_t$. In addition, such dependence is homogeneous over time. Under CA, SRA and MA, it further implies that the process $\{(\boldsymbol{S}_t, \boldsymbol{A}_t) : t \ge 0\}$ forms a time-homogeneous Markov chain. Let $p(b, \cdot)$ be the density function of the stationary distribution of $\{\boldsymbol{S}_t : t \ge 0\}$. Similarly, for a given non-dynamic policy $\boldsymbol{\pi}$, let $p(\boldsymbol{\pi}, \cdot)$ be the stationary density function of $\{\boldsymbol{S}_t : t \ge 0\}$ had all agents followed $\boldsymbol{\pi}$. When (A1) holds and the initial distribution of $\{\boldsymbol{S}_t : t \ge 0\}$ equals its stationary distribution, the stationarity condition in (A2) is automatically satisfied. The second part of (A2) holds when $\{(\boldsymbol{S}_t, \boldsymbol{A}_t) : t \ge 0\}$ satisfies geometric ergodicity [see Theorem 3.7 of 6]. Geometric ergodicity is weaker than the uniform ergodicity condition imposed in the existing reinforcement learning literature [2, 31].

IS based estimator. In the following, we first consider a potential estimator for $V(\boldsymbol{\pi})$, which is built on the value estimator proposed by [15] in a single-agent system. We then discuss its limitation and present our IS based estimator. Note that $V_i(\boldsymbol{\pi}) = \int_{\mathbb{S}} r_i(\boldsymbol{\pi}, \boldsymbol{s})p(\boldsymbol{\pi}, \boldsymbol{s})d\boldsymbol{s}$. Let $\omega(\boldsymbol{\pi}, \boldsymbol{s}) = p(b, \boldsymbol{s})/p(\boldsymbol{\pi}, \boldsymbol{s})$. By the change-of-measure equality, we obtain

$$V_i(\boldsymbol{\pi}) = \int_{\mathbb{S}_i} \omega(\boldsymbol{\pi}, \boldsymbol{s})r_i(\boldsymbol{\pi}, \boldsymbol{s})p(b, \boldsymbol{s})d\boldsymbol{s} = \mathbb{E}\omega(\boldsymbol{\pi}, \boldsymbol{S}_t)r_i(\boldsymbol{\pi}, \boldsymbol{S}_t) = \mathbb{E}\omega(\boldsymbol{\pi}, \boldsymbol{S}_t)\frac{\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi})}{b(\boldsymbol{\pi}|\boldsymbol{S}_t)}R_{i,t}. \quad (2)$$

A natural estimator for $V_i(\boldsymbol{\pi})$ is the following IS based estimator $T^{-1}\sum_{t=0}^{T-1}\widehat{\omega}(\boldsymbol{\pi}, \boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi})R_{i,t}/b(\boldsymbol{\pi}|\boldsymbol{S}_t)$, for some estimated $\widehat{\omega}$. The corresponding estimator for $V(\boldsymbol{\pi})$ is given by $(NT)^{-1}\sum_{i=1}^N\sum_{t=0}^T \widehat{\omega}(\boldsymbol{\pi}, \boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi})R_{i,t}/b(\boldsymbol{\pi}|\boldsymbol{S}_t)$.

In a multi-agent system, the above estimator has two limitations. The first is that it suffers from the high variance introduced by the importance ratio $\omega(\boldsymbol{\pi}, \boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi})/b(\boldsymbol{\pi}|\boldsymbol{S}_t)$. To better illustrate this, suppose the state-action pairs are independent across different agents. Then the overall ratio is the product of ratios associated with each single agent. As such, variances in each individual ratio accumulate multiplicatively, so the overall ratio can have an extremely high variance for large N. The second is that consistent estimation of $\omega(\boldsymbol{\pi}, \boldsymbol{S}_t)$ is extremely challenging with high-dimensional state-action space and limited observations. One naive approach is to replace the overall weight in (2) by the individual ratio associated with the i-th agent. However, such an approach ignores the interference between different spatial units, leading to a biased value estimator.

To address the these concerns, we consider factorizing the ratio based on the mean-field approximation. To this end, for any $1 \le i \le N$, let $\mathcal{N}(i)$ denote the index set of the neighboring agents of agent i. Let $\widetilde{S}_i$ and $\widetilde{A}_i$ be some mean-field functions of the local states and actions related to the i-th agent, respectively. For instance, one might set $\widetilde{S}_i$ and $\widetilde{A}_i$ to the average state and action over its neighbors,

$$\widetilde{S}_i(\boldsymbol{s}) = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} s_j \text{ and } \widetilde{A}_i(\boldsymbol{a}) = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} a_j, \quad \forall i,$$

where $|\mathcal{N}(i)|$ denotes the number of candidates in $\mathcal{N}(i)$ and (s_i, a_i) corresponds to the state-action pair associated with the i-th agent. For any random vectors Z_1, Z_2, Z_3, we use the notation $Z_1 \perp\!\!\!\perp Z_2 | Z_3$ to indicate that Z_1 and Z_2 are independent conditional on Z_3.

(A3)(i) For each $i \in \{1, \cdots, N\}$, there exists some function r_i^* such that for any $\boldsymbol{s} \in \mathbb{S}$, $\boldsymbol{a} \in \{0,1\}^N$, we have $r_i(\boldsymbol{a}, \boldsymbol{s}) = r_i^*(a_i, \widetilde{A}_i(\boldsymbol{a}), s_0, s_i, \widetilde{S}_i(\boldsymbol{s}))$ where s_0 denotes the sub-vector of $\boldsymbol{s}$ associated with the global state; (ii) $S_{i,t+1} \perp\!\!\!\perp \boldsymbol{S}_t, \boldsymbol{A}_t | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}$ for any i, where we use a shorthand and write $\widetilde{S}_i(S_{i,t}) = \widetilde{S}_{i,t}$, $\widetilde{A}_i(A_{i,t}) = \widetilde{A}_{i,t}$; (iii) $S_{0,t+1} \perp\!\!\!\perp \boldsymbol{S}_t, \boldsymbol{A}_t | S_{0,t}$; (iv) $\widetilde{S}_{i,t+1} \perp\!\!\!\perp \boldsymbol{S}_t, \boldsymbol{A}_t | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}$ for any i.

The first two parts of (A3) requires the immediate reward and future state in each spatial unit to depend on the current state-action pairs only through its neighbors'. It is generally conceived that these conditions hold in our applications. Specifically, the state-action pair at one area can affect the outcome of other locations only through its impact on the distribution of drivers. Within each time unit, each driver can travel at most from one spatial unit to its neighbor. Hence, the distribution of drivers in one location is independent of the state-action pairs in nonadjacent areas. The third part of (A3) requires the transition dynamics of the global state to be independent of region-specific state-action pairs. This condition automatically holds when the global state corresponds to some deterministic variables such as the time of day.

(A3)(ii)-(iv) together with (A1) implies that the process $\{(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) : t \geq 0\}$ satisfies the Markov property. Let $p_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ and $p_i(b, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ denote the density function of the stationary distribution of $\{(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) : t \geq 0\}$ under $\boldsymbol{\pi}$ and b, respectively. Let $\omega_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) = p_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) / p_i(b, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$. Using similar arguments in (2), we have by (A3)(i) that

$$\begin{aligned} V_i(\boldsymbol{\pi}) &= \int_{s_0, s_i, \tilde{s}_i} \omega_i(\boldsymbol{\pi}, s_0, s_i, \tilde{s}_i) r_i^*(\boldsymbol{\pi}, s_0, s_i, \tilde{s}_i) p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i \\ &= \mathbb{E}\omega_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \tilde{S}_{i,t}) \mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi})) R_{i,t} / b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}), \end{aligned} \tag{3}$$

where $b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ denotes treatment assignment probability $\mathbb{P}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}) | S_{0,t}, S_{i,t}, \tilde{S}_{i,t})$. In experiments where $\boldsymbol{A}_t$ is independent of $\boldsymbol{S}_t$, b_i can be explicitly calculated. Otherwise, b_i can be estimated by the state-of-the-art machine learning algorithms (see Appendix A.3 in the supplementary article for details).

Motivated by (3), we consider the following IS based estimator,

$$\widehat{V}_i^{\mathrm{IS}}(\boldsymbol{\pi}) = \frac{1}{T} \sum_{t=0}^{T-1} \widehat{\omega}_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} R_{i,t},$$

for some estimated $\widehat{\omega}_i$. Since the sampling ratio in $\widehat{V}_i^{\mathrm{IS}}(\boldsymbol{\pi})$ is a function of $(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t})$ only, $\widehat{V}_i^{\mathrm{IS}}(\boldsymbol{\pi})$ has a much smaller variance compared to the value estimator outlined at the beginning of this section. It remains to estimate ω_i. Our procedure is motivated by the following lemma.

Lemma 1 *Under (A1) and (A3)(ii)-(iv), we have* $\mathbb{E}\Delta_{i,t}(\omega_i) f(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) = 0$ *for any* i, t *and function* f *where*

$$\Delta_{i,t}(\omega_i) = \omega_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - \omega_i(\boldsymbol{\pi}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}).$$

Lemma 1 motivates us to compute $\widehat{\omega}_i$ by minimizing the following loss function,

$$\widehat{\omega}_i = \arg\min_{\omega_i \in \Omega} \sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} \Delta_{i,t}(\omega_i) f(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) \right|^2, \tag{4}$$

for some function classes Ω and $\mathcal{F}$. In our implementation, we set Ω to a neural network class and $\mathcal{F}$ to a ball of a reproducing kernel Hilbert space (RKHS). Additional details of the algorithm are given in Appendix A.1 of the supplementary article to save space.

Given $\widehat{V}_i^{\mathrm{IS}}(\boldsymbol{\pi})$, the corresponding estimator for the average value $V(\boldsymbol{\pi})$ is given by $\widehat{V}^{\mathrm{IS}}(\boldsymbol{\pi}) = N^{-1} \sum_{i=1}^N \widehat{V}_i^{\mathrm{IS}}(\boldsymbol{\pi})$.

Doubly-robust estimator. Compared to $\widehat{V}^{\mathrm{IS}}(\boldsymbol{\pi})$, the doubly-robust (DR) estimator offers protection against model misspecification of the density ratio and is more efficient in general. Before presenting the estimator, we introduce some notations.

Under a given policy $\boldsymbol{\pi}$, define the Q-function associated with the i-th agent as

$$Q_i(\boldsymbol{\pi}; \boldsymbol{a}, \boldsymbol{s}) = \sum_{t=0}^{+\infty} \mathbb{E}[\{R_{i,t}^*(\boldsymbol{\pi}(\boldsymbol{a})) - V_i(\boldsymbol{\pi})\} | \boldsymbol{S}_0 = \boldsymbol{s}], \quad \forall \boldsymbol{s} \in \mathbb{S}_0, \boldsymbol{a} \in \{0,1\}^N,$$

where $R_{i,t}^*(\boldsymbol{\pi}(\boldsymbol{a}))$ denotes the potential outcome in the i-th spatial unit that would occur at time t were the initial treatment equal to $\boldsymbol{a}$ and all other actions assigned according to $\boldsymbol{\pi}$. Different from the existing literature on reinforcement learning, we define the Q-function through potential outcomes rather than the observed data. We next derive a version of the Bellman equation for Q_i.

Lemma 2 $\mathbb{E}\{R_{i,t} + Q_i(\boldsymbol{\pi}; \boldsymbol{\pi}, \boldsymbol{S}_{t+1}) | \boldsymbol{S}_t, \boldsymbol{A}_t\} = V_i(\boldsymbol{\pi}) + Q_i(\boldsymbol{\pi}; \boldsymbol{A}_t, \boldsymbol{S}_t)$ *almost surely for any* i, t.

The DR estimator for $V_i(\boldsymbol{\pi})$ takes the following form,

$$\widetilde{V}_i(\boldsymbol{\pi}) + \frac{1}{T}\sum_{t=0}^{T-1} \widetilde{\omega}(\boldsymbol{\pi}, \boldsymbol{S}_t) \frac{\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi})}{b(\boldsymbol{\pi}|\boldsymbol{S}_t)} \{R_{i,t} + \widetilde{Q}_i(\boldsymbol{\pi}, \boldsymbol{S}_{t+1}) - \widetilde{Q}_i(\boldsymbol{A}_t, \boldsymbol{S}_t) - \widetilde{V}_i(\boldsymbol{\pi})\}, \tag{5}$$

where $\widetilde{V}_i(\boldsymbol{\pi})$ denotes some initial estimator for $V_i(\boldsymbol{\pi})$, $\widetilde{\omega}$ and $\widetilde{Q}_i$ stand for estimators for ω and Q_i, respectively. Note that by Lemma 2, the second term in (5) has zero mean when $(\widetilde{Q}_i, \widetilde{V}_i(\boldsymbol{\pi})) = (Q_i, V_i(\boldsymbol{\pi}))$. When $\widetilde{\omega} = \omega$, (5) is equivalent to the IS-based estimator $(T+1)^{-1}\sum_{t=0}^{T} \omega(\boldsymbol{\pi}, \boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}) R_{i,t} b^{-1}(\boldsymbol{\pi}|\boldsymbol{S}_t)$. Based on the above discussion, one can verify that (5) is consistent when either $\widetilde{\omega} = \omega$ or $(\widetilde{Q}_i, \widetilde{V}_i(\boldsymbol{\pi})) = (Q_i, V_i(\boldsymbol{\pi}))$.

However, due to the presence of high-dimensional state-action space, the estimator outlined in (5) suffers from high variance. In addition, consistent estimation of ω and Q_i are extremely difficult. To address these concerns, we replace the density ratio in (5) by $\widehat{\omega}_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))/b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ where the estimator $\widehat{\omega}_i$ is defined in (4). To enable consistent estimation of Q_i, we consider factorizing Q_i based on mean-field approximation as well. To this end, we introduce the following condition.

(A4) For each $i \in \{1, \cdots, N\}$, there exists some function Q_i^* such that for any $s \in \mathbb{S}$, $\boldsymbol{a} \in \{0,1\}^N$, we have $Q_i(\boldsymbol{\pi}; \boldsymbol{a}, \boldsymbol{s}) = Q_i^*(\boldsymbol{\pi}; a_i, \widetilde{A}_i(\boldsymbol{a}), s_0, s_i, \widetilde{S}_i(\boldsymbol{s}))$.

To learn Q_i^* and $V_i(\boldsymbol{\pi})$, we extend the regularized policy iteration algorithm [9, 14] to our setup. The key ingredient of the algorithm lies in minimizing a regularized version of the Bellman residual to work with rich nonparametric function class and controls its complexity. In our implementation, we use RKHS as the function class to approximate the Q-function. Detailed procedure is given in Appendix A.2 of the supplementary article to save space. Let $\widehat{Q}_i$ and $\widehat{V}_i(\boldsymbol{\pi})$ be the corresponding estimator, we define our value estimator by

$$\widehat{V}_i^{\mathrm{DR}}(\boldsymbol{\pi}) = \widehat{V}_i(\boldsymbol{\pi}) + \frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} \{R_{i,t} + \widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t} - \widehat{V}_i(\boldsymbol{\pi})\},$$

where $\widehat{Q}_{i,t+1}(\boldsymbol{\pi})$, $\widehat{Q}_{i,t}$ and $\widehat{\omega}_{i,t}$ are shorthand for $\widehat{Q}_i(\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})$, $\widehat{Q}_i(A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ and $\widehat{\omega}_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$, respectively. The corresponding estimator for $V(\boldsymbol{\pi})$ is given by $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi}) = N^{-1}\sum_{i=1}^{N} \widehat{V}_i^{\mathrm{DR}}(\boldsymbol{\pi})$.

Let $V_i^*(\boldsymbol{\pi})$ and ω_i^* be the population limit of $\widehat{V}_i(\boldsymbol{\pi})$ and $\widehat{\omega}_i$. We require $V_i^*(\boldsymbol{\pi}) = V_i(\boldsymbol{\pi})$ when (A4) holds and $\omega_i^* = \omega_i$ when (A3) holds. To better understand our theoretical results, we begin by investigating the performance of an "oracle" estimator $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$ which works as if the true values Q_i^*, ω_i^* and $V_i^*(\boldsymbol{\pi})$ were known. Specifically, let

$$\widehat{V}_i^{\mathrm{DR}*}(\boldsymbol{\pi}) = V_i^*(\boldsymbol{\pi}) + \frac{1}{T}\sum_{t=0}^{T-1} \omega_{i,t}^* \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} \{R_{i,t} + Q_{i,t+1}^*(\boldsymbol{\pi}) - Q_{i,t}^* - V_i^*(\boldsymbol{\pi})\},$$

where $Q_{i,t+1}^*(\boldsymbol{\pi})$, $Q_{i,t}^*$ and $\omega_{i,t}^*$ are shorthand for $Q_i^*(\boldsymbol{\pi}; \pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})$, $Q_i^*(\boldsymbol{\pi}; \pi_i, \widetilde{A}_{i,t+1}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})$ and $\omega_i^*(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$. The oracle estimator is given by $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi}) = N^{-1}\sum_{i=1}^{N} \widehat{V}_i^{\mathrm{DR}*}(\boldsymbol{\pi})$.

In Theorem 1, we establish the doubly-robustness property of the oracle estimator. Specifically, we show the oracle estimator is $(NT)^{-1/2}$-consistent and asymptotically normal when one of the mean-field approximation is valid.

Theorem 1 *Suppose (A1) and (A2) hold, $NTVar\{\widehat{V}^{DR*}(\boldsymbol{\pi})\} \to \sigma^2 > 0$ and $T \to \infty$. Suppose $\{R_{i,t}, Q_i^*, \omega_i, V_i(\boldsymbol{\pi}) : 1 \leq i \leq N, t \geq 0\}$ are uniformly bounded from infinity, the set of functions $\{b_i : 1 \leq i \leq N\}$ are uniformly bounded from zero. Then as either (A3) or (A4) holds, we have*

$$\sqrt{NT}\{\widehat{V}^{DR*}(\boldsymbol{\pi}) - V(\boldsymbol{\pi})\} \xrightarrow{d} N(0, \sigma^2).$$

We next investigate the statistical properties of the proposed estimator $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$. We need some technical conditions on the estimated density ratio and Q-function. To save space, we summarize these conditions in (A5), (A6) and present them in Appendix B. Theorem 2 establishes the doubly-robustness property of our estimator.

Theorem 2 (doubly-robustness) *Suppose the conditions in Theorem 1 hold. Suppose (A5) holds. Then as either (A3) or (A4) holds, we have $\widehat{V}^{DR*}(\boldsymbol{\pi}) - V(\boldsymbol{\pi}) = o_p(1)$.*

In Theorem 3, we show our value estimator achieves the "oracle" property when both mean-field approximations are valid. Specifically, it is $(NT)^{-1/2}$-consistent and asymptotically normal with the asymptotic variance equal to that of the oracle estimator.

Theorem 3 (oracle property) *Suppose the conditions in Theorem 2 hold. Suppose (A6) holds. Then when both (A3) and (A4) hold, we have $\sqrt{NT}\{\widehat{V}^{DR}(\boldsymbol{\pi}) - V(\boldsymbol{\pi})\} \xrightarrow{d} N(0, \sigma^2)$.*

4 Numerical experiments

4.1 Synthetic data

In this section, we conduct a simulation experiment which mimics our motivating ride-sharing example. Specifically, we consider a 5 by 5 gird world with regions indexed as $i \in \{1, \ldots, 25\}$. During $(t-1, t]$, the company records the number of drivers $D_{i,t}$ and orders $O_{i,t}$ in each region i. The degree of mismatch is measured by $M_{t,g} = 0.5 * (1 - \frac{|D_{t,g} - O_{t,g}|}{|1 + D_{t,g} + O_{t,g}|}) + 0.5 * M_{t-1,g}$. Define the state as $S_{i,t} = (O_{i,t}, D_{i,t}, M_{i,t})^T$. Then at time t, the company decides the subsidizing policies $\{A_{i,t}\}_{i=1}^{25}$ for $(t, t+1]$. The reward is defined as $R_{t,i} = M_{t+1,i} min(D_{t+1,i}, O_{t+1,i}) + e_{t,i}^R$, where $e_{t,i}^R$ follows $\mathcal{N}(0, \sigma_R^2)$. We sample $O_{i,t}$ from $Poisson(u_i^O)$, where $\{u_i^O\}_{i=1}^{25}$ represent a fixed spatial pattern and are randomly generated from $logNormal(4.6, 0.3)$. Drivers will be attracted to a region by both the subsidizing policy and the availability of orders. To characterize this, we first assign an attraction parameter to each region as $u_{i,t} = exp(A_{i,t}) + 0.5(\frac{O_{i,t}}{1+D_{i,t}})$, and then model the dynamics of driver as $D_{i,t+1} = \sum_{i \in \mathcal{N}(i)} (\frac{u_{i,t}}{\sum_{j \in \mathcal{N}(i)} u_{j,t}} D_{i,t})$. In each replication, after a burn-in period of length 50, a dataset of length $T = 336$ is generated following the behaviour policy $A_{i,t} \sim Bernoulli(0.5)$ for every i and t. Motivated by the business application, we focus on a family of threshold-based target policies $\{\boldsymbol{\pi}_c = (I_{\{u_1^O \geq c\}}, \ldots, I_{\{u_{25}^O \geq c\}}), c \geq 0\}$. The value of each $\boldsymbol{\pi}_c$ is estimated using this dataset, and the mean squared error (MSE) among 100 replications are recorded with the true value obtained via Monte Carlo.

Four related competing methods are considered: IS with mean-field approximation $\widehat{V}^{\mathrm{IS}}(\boldsymbol{\pi})$, DR without mean-field approximation introduced in (5), DR ignoring the spatial interference, and naive average of rewards in the observed data. The results for $c \in \{95, 100, 105, 110\}$ and different σ_R are presented in Figure 2. The spatial pattern of orders and target policies can be found in the supplementary article. It can be seen clearly that the proposed DR estimator consistently shows smaller MSEs than other methods, while DR without the spatial information or without mean-field approximation does not work well.

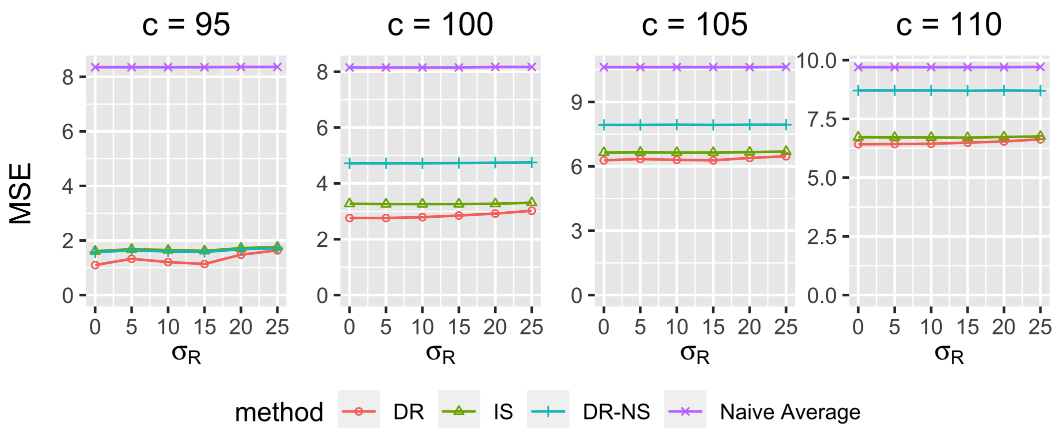

Figure 2: Off-policy evaluation results for the simulated ride-sharing subsidizing example. DR without spatial information is abbreviated as DR-NS. MSEs for DR without mean-field approximation are all larger than 10^4 and hence not plotted.

References

[1] Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *J. Amer. Statist. Assoc.*, 113(521):230–240, 2018.

[2] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

[3] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2019. NIH Public Access, 2019.

[4] Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *J. Amer. Statist. Assoc.*, 114(528):1665–1682, 2019.

[5] Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A. Murphy. Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.*, 113(523):1112–1121, 2018.

[6] Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144, 2005. Update of, and a supplement to, the 1986 original.

[7] Walter Dempsey, Peng Liao, Santosh Kumar, and Susan A Murphy. The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *arXiv preprint arXiv:1711.03587*, 2017.

[8] Ashkan Ertefaie. Constructing dynamic treatment regimes in infinite-horizon settings. *arXiv preprint arXiv:1406.0764*, 2014.

[9] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.*, 17:Paper No. 139, 66, 2016.

[10] Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *J. Amer. Statist. Assoc.*, 103(482):832–842, 2008.

[11] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661, 2016.

[12] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

[13] Eric B. Laber, Nick J. Meyer, Brian J. Reich, Krishna Pacifici, Jaime A. Collazo, and John M. Drake. Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 67(4):743–789, 2018.

[14] Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *arXiv preprint arXiv:1912.13088*, 2019.

[15] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

[16] Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, accepted, 2019.

[17] S. A. Murphy. Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(2):331–366, 2003.

[18] Bo Ning, Subhashis Ghosal, Jewell Thomas, et al. Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis*, 14(1):1–28, 2019.

[19] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*, pages 441–470. Springer, 2012.

[20] James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

[21] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[22] Donald B Rubin. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[23] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

[24] Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On causal inference in the presence of interference. *Stat. Methods Med. Res.*, 21(1):55–75, 2012.

[25] Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[26] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497, 2013.

[27] Masatoshi Uehara and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

[28] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*, 2018.

[29] Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

[30] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[31] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675, 2019.