THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Shi, Chengchun , Wan, Runzhe, Song, Ge, Luo, Shikai, Zhu, Hongtu & Song, Rui (2023) A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *Annals of Applied Statistics, 17*(4), 2701 - 2722. https://doi.org/10.1214/22-AOAS1700

https://researchonline.lse.ac.uk/id/eprint/117174/

Version: Accepted Version Licence: Creative Commons: Attribution 4.0

This document is the author's accepted version of the journal article. There may be differences between this version and the published version.

This author accepted manuscript has been made open access under the LSE open access publications policy and distributed under a Creative Commons: Attribution 4.0 license

LSE Research Online is the repository for research produced by the London School of Economics and Political Science. For more information, please refer to our Policies page or contact lseresearchonline@lse.ac.uk

A MULTI-AGENT REINFORCEMENT LEARNING FRAMEWORK FOR OFF-POLICY EVALUATION IN TWO-SIDED MARKETS

BY CHENGCHUN SHI[1] RUNZHE WAN[2] GE SONG[3] SHIKAI LUO[3] HONGTU ZHU[4] RUI SONG[2]

[1]*London School of Economics and Political Science, c.shi7@lse.ac.uk*

[2]*North Carolina State University, rwan@ncsu.edu; rsong@ncsu.edu*

[3]*Didi Chuxing, songge@didiglobal.com; luoshikai@didiglobal.com*

[4]*The Univeristy of North Carolina at Chapell Hill, htzhu@email.unc.edu*

The two-sided markets such as ride-sharing companies often involve a group of subjects who are making sequential decisions across time and/or location. With the rapid development of smart phones and internet of things, they have substantially transformed the transportation landscape of human beings. In this paper we consider large-scale fleet management in ride-sharing companies that involve multiple units in different areas receiving sequences of products (or treatments) over time. Major technical challenges, such as policy evaluation, arise in those studies because (i) spatial and temporal proximities induce interference between locations and times; and (ii) the large number of locations results in the curse of dimensionality. To address both challenges simultaneously, we introduce a multi-agent reinforcement learning (MARL) framework for carrying policy evaluation in these studies. We propose novel estimators for mean outcomes under different products that are consistent despite the high-dimensionality of state-action space. The proposed estimator works favorably in simulation experiments. We further illustrate our method using a real dataset obtained from a two-sided marketplace company to evaluate the effects of applying different subsidizing policies. A Python implementation of our proposed method is available at `https://github.com/RunzheStat/CausalMARL`.

1. Introduction. This paper concerns the applications in the two-sided markets that involve a group of subjects who are making sequential decisions across time and/or location. In particular, we consider large-scale fleet management in ride-sharing companies, such as Uber, Lyft and Didi. These companies form a typical two-sided market that enables efficient interactions between passengers and drivers (Armstrong, 2006; Rysman, 2009). With the rapid development of smart phones and internet of things, they have substantially transformed the transportation landscape of human beings (Frenken and Schor, 2017; Jin et al., 2018; Hagiu and Wright, 2019). With rich information on passenger demand and locations of taxi drivers, they significantly reduce taxi cruise time and passenger waiting time in comparison to traditional taxi systems (Li et al., 2011; Zhang et al., 2014; Miao et al., 2016). We use the numbers of drivers and call orders to measure the supply and demand at a given time and location. Both supply and demand are spatio-temporal processes and they interact with each other. These processes depend strongly on the platform's policies, and have a huge impact on the platform's outcomes of interest, such as drivers' income level and working time, passengers' satisfaction rate, order answering rate and order finishing rate, etc.

Keywords and phrases: Multi-Agent system, Reinforcement learning, Spatiotemporal studies, Policy evaluation.

A fundamental question of interest that we consider here is how to establish causal relationships between platform policies and platform's outcomes. In particular, we are interested in evaluating the causal effects of applying different subsidizing policies or recommendation programs to drivers or passengers in different spatial locations of a city. The purpose of implementing these policies is to balance the taxi supply and passenger demand across different areas of the city, so as to meet more passengers' requests and reduce drivers' vacant time. As an example, suppose a passenger opens the ride-sharing application on their smart phone and enters their destination. The platform will decide whether to recommend the customer to join a program and send them a coupon to discount this ride, depending on their locations. Such a recommendation increases the chance that the customer orders this particular ride to reduce the local drivers' vacant time. As another example, suppose the ride-sharing platform provides subsides to drivers in areas where there are more passenger call orders than the number of drivers. Implementing such a policy will attract more drivers to these areas, meeting more passengers' requests.

Solving this fundamental question faces at least two major challenges. The first one is that the spatial and temporal proximities in the aforementioned applications will induce interference between locations and times. As such, the outcome associated with each unit might depend on the treatments of all regions. Learning each spatial unit's value based on its own data only would yield a biased estimator. See e.g., the performance of the baseline estimator DR-NS in Sections 4 and 5. We notice that most of the existing policy evaluation methods in the literature focus on the setting where no interference occurs, that is, the outcome of each experimental unit is unaffected by treatment assignment of other units (see e.g., Zhang et al., 2012; Chakraborty et al., 2014; Dudík et al., 2014; Matsouaka et al., 2014; Luedtke and Van Der Laan, 2016; Belloni et al., 2017; Wu and Wang, 2020; Shi et al., 2020a). Such a no interference assumption is often referred to as the stable unit treatment value assumption (SUTVA, Rubin, 1980, 1986) in the causal inference literature. To elaborate the violation of SUTVA, let us revisit the example of applying driver-side subsidizing policies. As we have commented, applying a subsidizing policy at one location would attract some drivers from its neighbouring areas to that location, so the subsidizing policy at one location could influence various outcomes of those neighbouring areas, inducing interference among spatial units. Moreover, the subsidizing policy at a given time would affect both current and future supplies, inducing interference across time. In the passenger recommendation program example, sending coupons to certain passengers not only increases the number of current call orders, but the chance that these passenger use the app more frequently in the future as well. As such, applying the recommendation program will increase the number of future call orders, inducing interference across time. Thus, interference across time and/or among spatial units leads to the violation of SUTVA in our applications.

The second challenge is that the large number of spatial units results in the curse of dimensionality. Specifically, to implement the subsidizing policy or the customer recommendation program, the whole city is divided into $N = 8$ disjoint spatial units, as shown in Figure 1. For each spatial unit, the company can decide whether to apply certain promotion strategy to this region or not. In the presence of spatial interference, the amount of data needed to provide reliable policy value estimates grows exponentially with N, resulting in the curse of dimensionality. It remains challenging to effectively model the high-dimensional system without additional assumptions.

The aim of this paper is to evaluate the treatment effects of multiple policies in the presence of spatiotemporal interference. There is a huge literature on causal inference, but most existing works assume SUTVA (see e.g., Hirano et al., 2003; Imbens and Rubin, 2015; Wager and Athey, 2018; Yao et al., 2020). There has been substantial interest in the development of causal inference under interference. Our work falls into an emerging research topic on space-

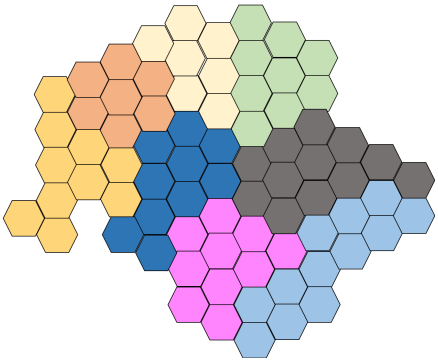

Fig 1: Visualization of $N = 8$ different spatial units in the city.

or time-dependent treatment effects evaluation (see e.g., Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012; Toulis and Kao, 2013; Halloran and Hudgens, 2016; Dempsey et al., 2017; Athey et al., 2018; Boruvka et al., 2018; Bhattacharya et al., 2019; Bojinov and Shephard, 2019; Ning et al., 2019; Reich et al., 2020). However, none of the above cited works studies the interference effects in both space and time. In particular, Reich et al. (2020) gave a systematic review of various statistical models for spatial causal inference and pinpoint some areas of future work. However, those models in Reich et al. (2020) were primarily motivated by research questions in environmental and epidemiological studies, so their generalization to two-sided markets remains unknown.

Reinforcement learning is a general machine learning technique that allows an agent to interact with a given environment, which has drawn more and more attention in the statistics literature, see Sutton and Barto (2018) for an overview. Recently, a number of proposals utilize reinforcement learning in mobile health or two-sided markets (Ertefaie, 2014; Luckett et al., 2019; Chen et al., 2020; Hu et al., 2019; Liao et al., 2020; Wang et al., 2021; Zhou et al., 2021; Li et al., 2022a,b; Liao et al., 2022; Shi et al., 2022a,b). In addition, there is a growing literature on adapting reinforcement learning to develop dynamic treatment regimes in precision medicine, to recommend treatment decisions based on individual patients' information (Murphy, 2003; Chakraborty et al., 2010; Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2013; Song et al., 2015; Zhao et al., 2015; Zhang et al., 2015, 2018; Zhu et al., 2017; Wang et al., 2018; Shi et al., 2018a,b; Mo et al., 2020; Meng et al., 2020; Cai et al., 2021; Fang et al., 2021). All these methods considered a single-agent setup where only one agent exists in the environment.

The proposed method is motivated by a line of research on multi-agent reinforcement learning (MARL) in the cooperative setting (see e.g., Zhang et al., 2019, for an overview). In MARL, multiple autonomous agents operate in a common environment. Each agent aims to maximize its own cumulative reward by interacting with the environment and other agents. Compared to the single-agent setup, MARL is much more challenging due to the presence of the high-dimensional action space induced by multiple agents. Most existing works on MARL focus on the *policy optimization* problem where each agent aims to identify an optimal policy that optimizes its long-term reward. In particular, Yang et al. (2018) developed a mean field Q-learning algorithm in the discounted-reward setting. We remark that many of these methods are not directly applicable to the problem of *policy evaluation*, where the objective is to learn the impact of a given policy using data collected possibly from a different behavior policy.

The contributions of this paper are summarized as follows. First, we introduce a multi-agent reinforcement learning framework for policy evaluation. Consider the example of ap-

plying driver-side subsidizing policies or passenger recommendation programs. 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 interference effects in space are modeled by the interactions between different agents, while the interference effects in time are modeled by the dynamic system transitions. See the causal diagram depicted in Figure 2 for an illustration. Estimation of the mean outcome under different products is reduced to the off-policy evaluation problem in MARL. To the best of our knowledge, this is the first work that explores MARL in the statistics literature. We remark that alternative to the MARL framework, the policy evaluation problem can be equivalently formulated using a single-agent setup whose action space is $\{0,1\}^N$ and reward is the sum of individual rewards of the agents. However, as mentioned earlier, it remains challenging to handle the exponentially large action space under the single-agent formulation. To the contrary, our formulation allows us to borrow ideas from the MARL literature to simultaneously address the challenges of dealing with spatiotemporal interference as well as the curse of dimensionality.

Second, we propose an original off-policy policy evaluation procedure in MARL. A number of off-policy evaluation algorithms have been developed under the single-agent setup (see e.g., Thomas et al., 2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Liu et al., 2018; Kallus and Uehara, 2019; Tang et al., 2019; Uehara et al., 2020; Shi et al., 2021; Chen and Qi, 2022). Directly applying these methods will yield value estimators with large variances, due to the curse of dimensionality. See the performance of the baseline estimator and DR-NM in Sections 4 and 5. Our proposal is the first to harness the power of modern MARL and single-agent policy evaluation methods, while establishing rigorous statistical guarantees, when tackling an important business question in two-sided markets. 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 both the density ratio and Q-function based on the mean-field approximation and aggregating these estimators properly to satisfy the doubly-robustness property in the average reward setting. Event though ideas such as the mean-field approximation and doubly robust estimation are not completely new, how to integrate them properly and effectively into a value estimator with desired theoretical guarantees is nontrivial, and is one of the main contributions of this article. In particular, we propose an original mean-field algorithm to approximate the density ratio in MARL. We also extend Yang et al. (2018)'s proposal to the average-reward setting, which is more suitable for our purpose of policy evaluation. 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 that 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.

Third, we systematically study how to test the mean-field approximation assumption from the observed data. Specifically, we show that under an additive noise model assumption, the mean-field approximation assumption is reduced to certain conditional mean independence assumptions. See Appendix C for details. This observation allows us to employ existing state-of-the-art conditional independence tests to examine the validity of mean-field approximation in practice. To our knowledge, this is the first time that a formal statistical test for such an assumption is proposed. We illustrate the idea using the forward-backward learning-based test developed by Shi et al. (2020b) in our numerical studies.

Finally, we 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). Our theory allows the number of spatial units, denoted as N, to be either bounded or diverge to infinity. Therefore,

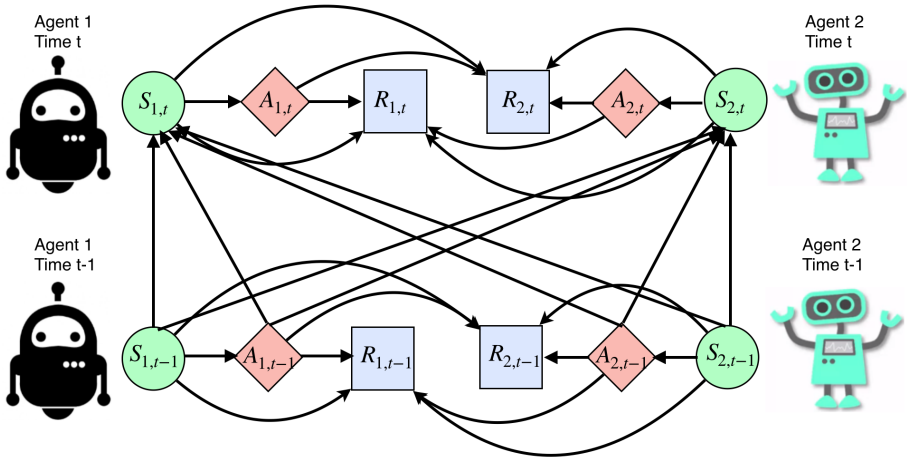

Fig 2: Causal diagram for a multi-agent system with two agents. $(S_{j,t}, A_{j,t}, R_{j,t})$ represents the state-treatment-outcome triplet of the j-th agent at time t.

the proposed estimator offers a useful policy evaluation tool to a wide range of applications in the presence of spatiotemporal interference. To prove these results, we develop an exponential inequality for the suprema of empirical processes under weak dependence (Lemma 3), which is useful for finite-sample analysis of machine learning estimates based on dependent observations. As we have mentioned earlier, most off-policy evaluation algorithms are developed under a single-agent setup. To our knowledge, these theoretical results have not been established in the existing literature.

The rest of the article is organized as follows. In Section 2, we describe the problem setup and introduce a potential outcome framework for MARL. We present our method in Section 3. Simulation studies are conducted in Section 4. In Section 5, we apply the proposed method to a dataset from a two-sided marketplace company to evaluate the effects of applying different subsidizing policies. Finally, we conclude our paper by a discussions section. Statistical property of our method is investigated in the appendix.

2. A potential outcome framework for MARL. In this section, we extend Rubin's potential outcome framework to the multi-agent system. This allows us to formulate our causal estimand. We first introduce some notations. For $1 \leq i \leq N$, we consider two treatments (actions) associated with the i-th spatial unit (agent) such that the action space is represented as $\mathcal{A}_i = \{0, 1\}$. Let $\mathbb{S}_i$ denote the state space associated with the i-th agent. In our application, the two treatments correspond to applying certain promotion strategy to a given spatial unit or not. The state variables include the number of call orders (demand) and available drivers (supply), and a supply-demand equilibrium metric that measures the degree of mismatch between orders and drivers within each region. The joint state and action spaces are given by $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2 \times \ldots \times \mathbb{S}_N$ and $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_N = \{0, 1\}^N$, respectively. For a sequence of N-dimensional vectors $\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_t \in \{0, 1\}^N$, we define a treatment history vector $\bar{\boldsymbol{a}}_t = (\boldsymbol{a}_0^\top, \boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_t^\top)^\top$ up to time t. For each $i \in \{1, \ldots, N\}$, let $S^*_{i,t+1}(\bar{\boldsymbol{a}}_t) \in \mathcal{S}_i$ and $R^*_{i,t}(\bar{\boldsymbol{a}}_t) \in \mathbb{R}$ be, respectively, the potential state and reward (outcome) associated with the i-th agent at time $t+1$ and time t, which would occur had all agents followed $\bar{\boldsymbol{a}}_t$. Moreover, different action histories would lead to different potential outcomes, and more importantly, these potential outcomes cannot be directly observed.

We first introduce a consistency assumption (CA) to link potential outcomes to the observed data. Let $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{1 \leq i \leq N, 0 \leq t < T}$ be the observed data, where

$(S_{i,t}, A_{i,t}, R_{i,t})$ 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 study. Let $\boldsymbol{A}_t = (A_{1,t}, \ldots, A_{N,t})^\top$ and $\bar{\boldsymbol{A}}_t = (\boldsymbol{A}_0^\top, \boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_t^\top)^\top$ be, respectively, the observed treatments at time t and until time t. The CA is given as follows.

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

Under CA, the potential outcomes are allowed to depend on not only past treatments, but also actions selected by other agents. The CA extends SUTVA to settings with spatiotemporal interference, since SUTVA requires $S_{i,t+1}^*$ and $R_{i,t}^*$ to be functions of $A_{i,t}$ only. We also remark that the notations $S_{i,t}^*$ and $R_{i,t}^*$ are used to denote potential outcomes. They are not deterministic functions of $\bar{\boldsymbol{A}}_{t-1}$ and $\bar{\boldsymbol{A}}_t$.

We next introduce a sequential randomization assumption (SRA) that guarantees the identifiability of our causal estimands.

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

We remark that SRA basically assumes that there is no unmeasured confounders. SRA is satisfied in randomized experiments, as in our real datasets. Under settings where SRA is violated, we can apply backdoor or frontdoor adjustment to the MARL setup to handle unmeasured confounders (see e.g., Wang et al., 2020). This is beyond the scope of the current paper and we leave it for future research. We also note that when $N = 1$, i.e., in the single agent setting, CA and SRA are commonly imposed in sequential decision making problems (see e.g., Murphy, 2003; Robins, 2004; Zhang et al., 2013; Ertefaie, 2014; Laber et al., 2018; Luckett et al., 2019).

Next, we introduce the Markov assumption (MA) and the conditional mean independence assumption (CMIA) to characterize the system transitions. These assumptions serve as the foundations of the existing state-of-the-art RL algorithms (see e.g., Sutton and Barto, 2018). Let $\boldsymbol{S}_t = (S_{1,t}, \cdots, S_{N,t})^\top$ and $\boldsymbol{R}_t = (R_{1,t}, \cdots, R_{N,t})^\top$.

(MA) There exists a Markov transition kernel $\mathcal{P} : \mathbb{S} \times \mathcal{A} \times \mathbb{S}$ 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 almost surely that

$$\Pr\{\boldsymbol{S}_{t+1} \in \mathcal{S} | \boldsymbol{A}_t, \boldsymbol{S}_t, \{\boldsymbol{A}_j, \boldsymbol{R}_j, \boldsymbol{S}_j\}_{0\le j<t}\} = \mathcal{P}(\mathcal{S}; \boldsymbol{A}_t, \boldsymbol{S}_t).$$

The transition kernel $\mathcal{P}$ characterizes the conditional distribution of the future state vector given the current state-action pair. MA assumes the system dynamics are homogeneous over time. This enables consistent estimation of our causal estimands. We remark that MA is testable from the observed data; see e.g., the goodness-of-fit test developed by Shi et al. (2020b).

(CMIA) There exist functions $r_1, \ldots, r_N$ such that for any $1 \le i \le N$, $t \ge 0$, $\bar{\boldsymbol{a}}_t \in \{0,1\}^{N(t+1)}$, we have $\mathbb{E}(R_{i,t} | \boldsymbol{A}_t, \boldsymbol{S}_t, \{\boldsymbol{A}_j, \boldsymbol{R}_j, \boldsymbol{S}_j\}_{0\le j<t}) = r_i(\boldsymbol{A}_t, \boldsymbol{S}_t)$ almost surely.

For each $1 \le i \le N$, r_i corresponds to the conditional mean function of $R_{i,t}$ given the state-action pair at time t. CMIA is similar to MA in the sense that it also imposes certain conditional independence assumption between the reward and the past data history. Meanwhile, it is weaker than requiring $R_{i,t}$ to be independent of $(A_j, R_j, S_j)_{0\le j<t}$ conditional on (A_t, S_t), since it only requires the conditional mean of $R_{i,t}$ to satisfy the independence assumption. It is also weaker than the condition that requires the reward to be a deterministic function of the state-action pair. The latter condition is commonly imposed in the literature (see e.g., Ertefaie, 2014; Luckett et al., 2019).

In practice, to ensure MA and CMIA to be satisfied, we can construct the state by concatenating measurements over multiple decision points till the Markov and conditional mean independence properties are satisfied. In addition, to guarantee the transition kernel $\mathcal{P}$ and the reward functions $\{r_i\}_i$ are time-homogeneous, we can include some auxiliary variables (e.g., time of the day) in the state. See our real data analysis in Section 5 for details. Without special saying, we assume that CA, SRA, MA and CMIA hold throughout this paper.

We next introduce the average treatment effect (ATE) for multi-agent systems below. We focus on the class of stationary policies indexed by some $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$, where each π_i is a binary-valued function of the current state vector. Under $\boldsymbol{\pi}$, the i-th spatial unit will receive the treatment $\pi_i(\boldsymbol{S}_t)$ at time t. As we have commented, applying a certain promotion strategy $\boldsymbol{\pi}$ in our application has both short-term and long-term benefits. We are thus interested in evaluating the average reward under $\boldsymbol{\pi}$. This allows the company to decide whether to apply such a dynamic policy in a given city or not, under some budget constraints. For any such policy $\boldsymbol{\pi}$, let $\bar{\boldsymbol{\pi}}_0 = \boldsymbol{\pi}(\boldsymbol{S}_0) = (\pi_1(\boldsymbol{S}_0), \ldots, \pi_N(\boldsymbol{S}_0))^\top$ be the initial action vector assigned according to $\boldsymbol{\pi}$. Then we recursively define $\bar{\boldsymbol{\pi}}_t = (\bar{\boldsymbol{\pi}}_{t-1}^\top, \boldsymbol{\pi}(S_t^*(\bar{\boldsymbol{\pi}}_{t-1})))^\top$ as the treatment assignment history under $\boldsymbol{\pi}$ up to time t for $t \geq 1$.

Our objective is to evaluate the long term value of a given policy $\boldsymbol{\pi}$, defined as

$$V(\boldsymbol{\pi}) = \lim_{t\to\infty} \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t} \mathbb{E} R_{i,j}^*(\bar{\boldsymbol{\pi}}_j), \tag{1}$$

where $R_{i,j}^*(\bar{\boldsymbol{\pi}}_j)$ denotes the potential outcomes of the i-th agent that would occur at time t had all agents followed the dynamic policy $\boldsymbol{\pi}$. We note that $V(\boldsymbol{\pi})$ can be represented by $N^{-1}\sum_{i=1}^N V_i(\boldsymbol{\pi})$, where $V_i(\boldsymbol{\pi}) = \lim_{t\to\infty} t^{-1} \sum_{j=0}^t \mathbb{E} R_{i,j}^*(\bar{\boldsymbol{\pi}}_j)$. Consequently, to evaluate $V(\boldsymbol{\pi})$, it suffices to estimate $V_i(\boldsymbol{\pi})$ for $i = 1, \cdots, N$.

3. Off-policy evaluation in MARL. In this section, we first propose an importance-sampling (IS) based estimator for $V_i(\boldsymbol{\pi})$ and then develop a doubly-robust version. We next detail some major steps in constructing these estimators.

3.1. *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 Liu et al. (2018) in a single-agent system. We then discuss its limitation and present our IS based estimator.

To detail the method, we assume the system follows a stationary behavior policy $b(\cdot)$ such that

$$\Pr(\boldsymbol{A}_t = \boldsymbol{a}_t | \{A_{i,j}, S_{i,j}, R_{i,j}\}_{1\leq i\leq N, 0\leq j<t} \cup \{S_{i,t}\}_{1\leq i\leq N}) = b(\boldsymbol{a}_t | \boldsymbol{S}_t),$$

holds for any $\boldsymbol{a}_t \in \{0,1\}^N$. In other words, $\boldsymbol{A}_t$ depends on past observations only through $\boldsymbol{S}_t$. It implies that the process $\{(\boldsymbol{S}_t, \boldsymbol{A}_t)\}_{t\geq 0}$ forms a time-homogeneous Markov chain. This assumption is satisfied in our application where the data are generated from a completely randomized experiment with $b(\boldsymbol{A}_t|\boldsymbol{S}_t) = 0.5^N$ for any t. Meanwhile, we also allow b to rely on the set of current state variables.

Let $p_b(\boldsymbol{s})$ be the density function of the stationary distribution of the stochastic process $\{\boldsymbol{S}_t\}_{t\geq 0}$. Similarly, for a given $\boldsymbol{\pi}$, let $p_{\boldsymbol{\pi}}(\boldsymbol{s})$ be the stationary density function of $\{\boldsymbol{S}_t\}_{t\geq 0}$ had all agents followed $\boldsymbol{\pi}$. When the process $\{\boldsymbol{S}_t\}_t$ reaches its stationary distribution, it follows from the change-of-measure theorem that

$$\begin{aligned} V_i(\boldsymbol{\pi}) &= \int_{\mathbb{S}_i} \omega(\boldsymbol{s}) r_i(\boldsymbol{\pi}(\boldsymbol{s}), \boldsymbol{s}) p_b(\boldsymbol{s}) d\boldsymbol{s} = \mathbb{E}\{\omega(\boldsymbol{S}_t) r_i(\boldsymbol{\pi}(\boldsymbol{S}_t), \boldsymbol{S}_t)\} \\ &= \mathbb{E}\left\{\omega(\boldsymbol{S}_t) \frac{\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{S}_t))}{b(\boldsymbol{\pi}(\boldsymbol{S}_t)|\boldsymbol{S}_t)} R_{i,t}\right\}, \end{aligned} \tag{2}$$

where $\omega(\boldsymbol{s}) = p_{\boldsymbol{\pi}}(\boldsymbol{s})/p_b(\boldsymbol{s})$ and $\mathbb{I}(\cdot)$ denotes the indicator function. Thus, a natural estimator for $V_i(\boldsymbol{\pi})$ is the IS based estimator $\widehat{V}_i^{IS_0}(\boldsymbol{\pi}) = T^{-1}\sum_{t=0}^{T-1}\widehat{\omega}(\boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{S}_t))R_{i,t}/b(\boldsymbol{\pi}(\boldsymbol{S}_t)|\boldsymbol{S}_t)$ for some estimated $\widehat{\omega}$, leading to the IS estimator for $V(\boldsymbol{\pi})$:

$$\widehat{V}^{IS_0}(\boldsymbol{\pi}) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\frac{\widehat{\omega}(\boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{S}_t))R_{i,t}}{b(\boldsymbol{\pi}(\boldsymbol{S}_t)|\boldsymbol{S}_t)}.$$

In a multi-agent system, the above estimator $\widehat{V}_i^{IS_0}(\boldsymbol{\pi})$ has two major limitations. The first one is that it suffers from high variance introduced by the importance ratio $\omega(\boldsymbol{S}_t)\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{S}_t))/b(\boldsymbol{\pi}(\boldsymbol{S}_t)|\boldsymbol{S}_t)$. To better illustrate this, suppose that the state-action pairs are independent across different agents. In this case, the overall ratio is the product of ratios associated with each single agent, so variances in each individual ratio accumulate multiplicatively. Thus, the overall ratio can have an extremely high variance for large N. The second one is that consistently estimation of $\omega(\cdot)$ is extremely challenging for high-dimensional state-action space and limited observations. One naive approach could replace the overall weight in (2) by the individual ratio associated with the i-th agent, but it would ignore the interference between different spatial units, leading to a biased value estimator.

To address these two limitations, we propose to factorize the importance ratio by using the mean-field approximation procedure. The key idea of this procedure is to approximate the reward function r_i as a function of the state-action pairs of the i-th agent and its neighbours only. This allows us to focus on the density ratio of these restricted state-action pairs as detailed in (5). As the input of such a density ratio is reduced to a moderate scale, the variance of the value estimator is dramatically reduced. It also enables consistent estimation of the density ratio.

We next detail the mean-field approximation procedure. For any $1 \le i \le N$, let $\mathcal{N}(i)$ denote the index set of the neighboring agents of agent i. Let m_i^s and m_i^a be some mean-field functions of the local states and actions related to the i-th agent, respectively. For instance, one may set these functions to some averaged state and action over its neighbors, i.e.,

$$m_i^s(\boldsymbol{s}) = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} s_j \quad \text{and} \quad m_i^a(\boldsymbol{a}) = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} a_j, \tag{3}$$

for any 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 each $i \in \{1, \ldots, N\}$ and any $\boldsymbol{s} \in \mathbb{S}$, $\boldsymbol{a} \in \{0,1\}^N$, we adopt the following mean-field approximation,

$$r_i(\boldsymbol{a}, \boldsymbol{s}) = \bar{r}_i(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s})), \tag{4}$$

for some function $\bar{r}_i$.

We make a few remarks. First, it is generally conceived that (4) holds in many applications, such as the ride-sharing platform. Specifically, the state-action pair at one location 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 location to its neighbouring locations. Hence, the distribution of drivers in one location is independent of the state-action pairs in non-adjacent locations.

Second, it is possible to test (4) for some given mean-field functions m_i^a and m_i^s. Specifically, notice that (4) essentially requires the conditional mean of $R_{i,t}$ to be independent of $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. When $R_{i,t}$ satisfies the additive noise model assumption, $R_{i,t} = f(S_{i,t+1}, A_{i,t}, S_{i,t}) + \varepsilon_{i,t}$, for some mean zero random error $\varepsilon_{i,t}$, (4) holds if $S_{i,t+1}$ is conditionally independent of $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. As such, existing state-of-the-art conditional independence tests can be applied to test this assumption. We discuss this further in Section 4.

Recall that $p_b(\cdot)$ and $p_{\boldsymbol{\pi}}(\cdot)$ are the stationary distribution of $\boldsymbol{S}_t$ under b and $\boldsymbol{\pi}$, respectively. Let $p_{i,\boldsymbol{\pi}}(\cdot)$ denote the corresponding marginal distribution of the triplet $\widetilde{S}_{i,t} = (m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. Similarly, we define $p_{i,b}(\cdot)$. Let $\omega_i(\widetilde{S}_{i,t})$ denote the density ratio $p_{i,\boldsymbol{\pi}}(\widetilde{S}_{i,t})/p_{i,b}(\widetilde{S}_{i,t})$. It follows from similar arguments in (2) that $V_i(\boldsymbol{\pi})$ equals

$$\begin{aligned}&\int_{\tilde{s}_i} \omega_i(\tilde{s}_i)\bar{r}_i(\pi_i(\boldsymbol{s}), \tilde{s}_i)p_{i,b}(\tilde{s}_i)d\tilde{s}_i \\ =&\mathbb{E}\omega_i(\widetilde{S}_{i,t})\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}R_{i,t},\end{aligned} \tag{5}$$

where $b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})$ denotes the conditional probability $\Pr(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t))|\widetilde{S}_{i,t})$. In settings where $\boldsymbol{A}_t$ is independent of $\boldsymbol{S}_t$, as in our application, b_i can be explicitly calculated. More generally, b_i can be estimated by the state-of-the-art machine learning algorithms (see Appendix B in the supplement for details).

Motivated by (5), we consider a new IS based estimator of $V_i(\boldsymbol{\pi})$ as follows:

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

for some estimated $\widehat{\omega}_i$. Since the sampling ratio in $\widehat{V}_i^{\text{IS}}(\boldsymbol{\pi})$ is a function of $\widetilde{S}_{i,t}$, $A_{i,t}$ and $m_i^a(\boldsymbol{A}_t)$ only, $\widehat{V}_i^{\text{IS}}(\boldsymbol{\pi})$ has a much smaller variance compared to the value estimator outlined at the beginning of this section. In addition, consistent estimation of ω_i is feasible since the dimension of the input of ω_i has been reduced to a moderate scale. Given $\widehat{V}_i^{\text{IS}}(\boldsymbol{\pi})$, the corresponding estimator for the average value $V(\boldsymbol{\pi})$ is given by $\widehat{V}^{\text{IS}}(\boldsymbol{\pi}) = N^{-1}\sum_{i=1}^N \widehat{V}_i^{\text{IS}}(\boldsymbol{\pi})$.

We discuss the estimating procedure for the density ratio ω_i in Section 3.3.

3.2. *Doubly-robust estimator.* Compared to $\widehat{V}^{\text{IS}}(\boldsymbol{\pi})$, the doubly-robust (DR) estimator offers protection against model misspecification of the density ratio and is more efficient in general. Kallus and Uehara (2019) developed a double reinforcement learning method for value evaluation in a single-agent discounted reward setting. In this section, we extend their proposal to a multi-agent average reward setup. Before presenting the estimator, we introduce the Q-function associated with the i-th agent under a given policy $\boldsymbol{\pi}$ as

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

where $\bar{\boldsymbol{\pi}}_t(\boldsymbol{a})$ denotes the treatment history up time t such that the initial treatment equal to $\boldsymbol{a}$ and all other actions assigned according to $\boldsymbol{\pi}$. We remark that $Q_i^{\boldsymbol{\pi}}(\boldsymbol{a}, \boldsymbol{s})$ is finite and well-defined when the Markov chain approaches its steady-state exponentially fast under $\boldsymbol{\pi}$. In that case, most of the differences on the right-hand-side will be earned in the first few iterations. Please see Section 8.2.1 of Puterman (1994) for details.

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

$$\widetilde{V}_i(\boldsymbol{\pi}) + \frac{1}{T}\sum_{t=0}^{T-1}\widetilde{\omega}(\boldsymbol{S}_t)\frac{\mathbb{I}(\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{S}_t))}{b(\boldsymbol{\pi}(\boldsymbol{S}_t)|\boldsymbol{S}_t)}\{R_{i,t} + \widetilde{Q}_i(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}), \boldsymbol{S}_{t+1}) - \widetilde{Q}_i(\boldsymbol{A}_t, \boldsymbol{S}_t) - \widetilde{V}_i(\boldsymbol{\pi})\}, \tag{6}$$

where $\widetilde{V}_i(\boldsymbol{\pi})$ denotes some initial estimator for $V_i(\boldsymbol{\pi})$ and $\widetilde{\omega}$ and $\widetilde{Q}_i$ stand for estimators for ω and Q_i, respectively. Note that by the Bellman equation (see Lemma 2 in Section 3.4 for details), the second term in (6) has zero mean when $(\widetilde{Q}_i, \widetilde{V}_i(\boldsymbol{\pi})) = (Q_i, V_i(\boldsymbol{\pi}))$. When

$\widetilde{\omega} = \omega$, it can be shown that (6) has the same asymptotic mean as the IS-based estimator. Based on the above discussion, one can verify that (6) 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 (6) 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 (6) by $\widehat{\omega}_i(\widetilde{S}_{i,t})\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))/b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})$. To enable consistent estimation of Q_i, we consider factorizing Q_i based on mean-field approximation as well. Specifically, for each $i \in \{1, \ldots, N\}$, and any $s \in \mathbb{S}$, $\boldsymbol{a} \in \{0,1\}^N$, we propose to approximate Q_i by

$$Q_i^{\boldsymbol{\pi}}(\boldsymbol{a}, \boldsymbol{s}) = \bar{Q}_i(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s})), \tag{7}$$

for some functions $\bar{Q}_i$.

When (4) holds, we show that (7) is satisfied if $(\pi(\boldsymbol{S}_{t+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S_{t+1}})), \widetilde{S}_{i,t+1})$ is conditionally independent of $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. See Appendix C for details. Similarly, existing state-of-the-art conditional independence tests can be applied to verify this assumption.

To learn $\bar{Q}_i$ and $V_i(\boldsymbol{\pi})$, we extend the regularized policy iteration algorithm (Farahmand et al., 2016; Liao et al., 2020) to our setup. The detailed procedure is given in Section 3.4. Meanwhile, other methods developed in single-agent systems (see e.g., Uehara et al., 2020) may be adopted as well. Let $\widehat{Q}_i$ and $\widehat{V}_i(\boldsymbol{\pi})$ denote the corresponding estimators, we define our value estimator

$$\begin{aligned} \widehat{V}_i^{\mathrm{DR}}(\boldsymbol{\pi}) = \widehat{V}_i(\boldsymbol{\pi}) + \frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_i(\widetilde{S}_{i,t}) \frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})} \\ \times\{R_{i,t} + \widehat{Q}_i(\pi_i(\boldsymbol{S}_{t+1}), \widetilde{S}_{i,t+1}) - \widehat{Q}_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t)) - \widehat{V}_i(\boldsymbol{\pi})\}. \end{aligned} \tag{8}$$

The corresponding estimator for $V(\boldsymbol{\pi})$ is given by $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi}) = N^{-1}\sum_{i=1}^N \widehat{V}_i^{\mathrm{DR}}(\boldsymbol{\pi})$.

To conclude this section, we present an overview of our theoretical results for $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$. Details are given in the appendix. Our theoretical studies are mostly concerned with an "oracle" estimator $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$, which works as if the true values $\bar{Q}_i$, ω_i^* and $V_i^*(\boldsymbol{\pi})$ were known. Specifically, let $\widehat{V}_i^{\mathrm{DR}*}(\boldsymbol{\pi})$ be a version of $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$ by replacing $\widehat{\omega}_i$, $\widehat{Q}_i$ and $\widehat{V}_i(\boldsymbol{\pi})$ in (8) with the corresponding population limits. 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, i.e., either (4) or (7) holds. In Theorem 2, we establish the doubly-robustness property of our estimator, i.e., $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$ is consistent when either (4) or (7) holds. In Theorem 3, we show our value estimator achieves the "oracle" property when both mean-field approximations (4) and (7) are valid. Specifically, it is $(NT)^{-1/2}$-consistent and asymptotically normal with the asymptotic variance equal to that of the oracle estimator. We remark that to establish these theoretical results, we require the state-action process $\{(\boldsymbol{S}_t, \boldsymbol{A}_t)\}$ to satisfy the exponential β-mixing condition; see (A1) in Appendix A. Under stationarity, this assumption is equivalent to require the underlying Markov chain to satisfy geometric ergodicity (see Theorem 3.7 of Bradley, 2005). It guarantees that the estimated values concentrates on their oracle values with high probability, allowing us to establish the oracle property. Please refer to the proof of Theorem 3 in Appendix D.3 for details.

3.3. *Estimation of the density ratio.* In the single-agent setup, there are multiple estimation methods available to produce an estimated density ratio (Liu et al., 2018; Kallus and

Uehara, 2019; Nachum et al., 2019; Liao et al., 2022). In our implementation, we extend the proposal in Liu et al. (2018) to the multi-agent setup. We first present an overview of the algorithm. A key observation is given by Lemma 1, which establishes the relationship between $\omega_i(\widetilde{S}_{i,t})$ and $\omega_i(\widetilde{S}_{i,t+1})$. Based on this lemma, the idea is to introduce a discriminator function to construct a mini-max loss function (see Equation (9)). Then ω_i is estimated by optimizing this loss function. We next present Lemma 1.

LEMMA 1. *Suppose $\widetilde{S}_{i,t+1}$ is independent of $(\boldsymbol{S}_t, \boldsymbol{A}_t)$ given $\widetilde{S}_{i,t}$, $A_{i,t}$ and $m_i^a(\boldsymbol{A}_t)$. We have $\mathbb{E}\Delta_{i,t}(\omega_i) f(\widetilde{S}_{i,t+1}) = 0$ for any i, t and function f where*

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

Under the conditions in Lemma 1, $\{(\widetilde{S}_{i,t}, A_{i,t}, m_i^a(\boldsymbol{A}_t))\}_{t\geq 0}$ forms a time-homogeneous Markov chain. This lemma motivates us to compute $\widehat{\omega}_i$ by minimizing the following loss function,

$$\widehat{\omega}_i = \arg\min_{\omega_i \in \Omega} \sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} \Delta_{i,t}(\omega_i) f(\widetilde{S}_{i,t+1}) \right|^2, \tag{9}$$

for some function classes Ω and $\mathcal{F}$. In our implementation, we set $\mathcal{F}$ to a unit ball of a reproducing kernel Hilbert space (RKHS), i.e.,

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} = 1\},$$

where

$$\mathcal{H} = \left\{ f(\cdot) = \sum_{t=0}^{T-1} b_t \kappa(\widetilde{S}_{i,t+1}; \cdot) : \{b_t\}_{t=0}^{T-1} \in \mathbb{R}^T \right\},$$

for some positive definite kernel $\kappa(\cdot;\cdot)$ and $\|\cdot\|_{\mathcal{H}}$ denotes the corresponding RKHS norm.

The use of RKHS enables us to derive a close-form expression for the objective function on the right-hand-side (RHS) of (9). Specifically, using similar arguments in the proof of Theorem 2 of Liu et al. (2018), the optimization problem in (9) is then reduced to

$$\widehat{\omega}_i = \arg\min_{\omega_i \in \Omega} \sum_{t_1=0}^{T-1} \sum_{t_2=0}^{T-1} \Delta_{i,t_1}(\omega_i) \Delta_{i,t_2}(\omega_i) \kappa(\widetilde{S}_{i,t_1+1}, \widetilde{S}_{i,t_2+1}).$$

It remains to specify the function class for Ω. Motivated by the approximation capabilities of neural networks, we set Ω to the class of multilayer perceptron networks. See Figure 3 for an illustration. We use different parameters to factorize different ω_i such that each $\widehat{\omega}_i$ is computed separately. Alternatively, one could allow different ω_i to share some common parameters. Stochastic gradient descent is applied to update the parameters in the neural network. We detail our procedure in Algorithm 1.

3.4. *Estimation of the Q-function and the value.* In this section, we describe methods to estimate compute $\widehat{Q}_i$ and the initial value estimator $\widehat{V}_i(\boldsymbol{\pi})$. The key ingredient of the algorithm lies in minimizing a regularized version of the Bellman residual to work with rich nonparametric function class, while simultaneously controlling its complexity. In our implementation, we use the RKHS as the function class to approximate the Q-function. A key observation is given by the following lemma,

Algorithm 1 Estimation of the density ratio.

Input: The data $\{(S_{i,j}, A_{i,j}, R_{i,j}) : 1 \le i \le N, 0 \le j < T\}$. A target policy $\boldsymbol{\pi}$.
Initialize: Initial the density ratio $\omega_i = \omega_{i,\theta}$ for $1 \le i \le N$, to be some neural networks parameterized by θ.
for iteration $= 1, 2, \cdots$ **do**

a Randomly sample a batch $\mathcal{M}$ from $\{0, 1, \cdots, T-1\}$.

b **Update** the parameter θ by $\theta \leftarrow \theta - \epsilon N^{-1} \sum_{i=1}^{N} \nabla_\theta D_i(\omega_{i,\theta}/z_{\omega_{i,\theta}})$ where $D_i(\omega_{i,\theta})$ is equal to

$$\frac{1}{|\mathcal{M}|} \sum_{t_1, t_2 \in \mathcal{M}} \Delta_{i,t_1}(\omega_{i,\theta}) \Delta_{i,t_2}(\omega_{i,\theta}) \kappa(\widetilde{S}_{i,t_1+1}, \widetilde{S}_{i,t_2+1}),$$

and $z_{\omega_{i,\theta}}$ is a normalization constant $z_{\omega_{i,\theta}} = |\mathcal{M}|^{-1} \sum_{t \in \mathcal{M}} \omega_{i,\theta}(\widetilde{S}_{i,t+1})$.

Output $\omega_{i,\theta}$ for $1 \le i \le N$.

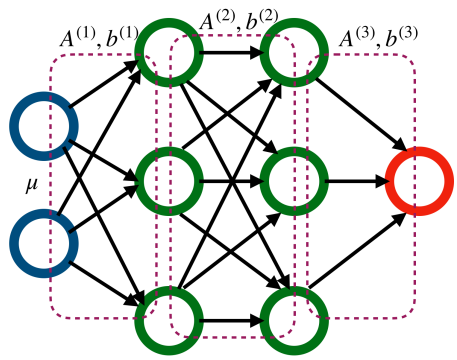

Fig 3: Illustration of multilayer perceptron with two hidden layers, $m_0 = 2$, $m_1 = m_2 = 3$. Here u is the input, $A^{(\ell)}$ and $b^{(\ell)}$ denote the corresponding parameters to produce the linear transformation for the $(\ell - 1)$th layer.

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

The equation in Lemma 2 is well-known as the Bellman equation in the average-reward case. Based on Lemma 2, it is tempting to estimate $Q_i^{\boldsymbol{\pi}}$ by directly minimizing

$$\arg\min_{(V_i, Q_i)} T^{-1} \sum_t \{R_{i,t} + Q_i(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}), \boldsymbol{S}_{t+1}) - V_i - Q_i(\boldsymbol{A}_t, \boldsymbol{S}_t)\}^2. \tag{10}$$

However, the resulting estimators are known to be biased when the Markov transition function is not deterministic (see e.g., Farahmand et al., 2016). To elaborate this, we note that the population limit of the objective function in (10) equals

$$\begin{aligned} &T^{-1} \sum_t \mathbb{E}[\mathbb{E}\{R_{i,t} + Q_i(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}), \boldsymbol{S}_{t+1}) | \boldsymbol{S}_t, \boldsymbol{A}_t\} - V_i - Q_i(\boldsymbol{A}_t, \boldsymbol{S}_t)]^2 \\ &\qquad + T^{-1} \sum_t \mathbb{E}[\mathrm{Var}\{R_{i,t} + Q_i(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}), \boldsymbol{S}_{t+1}) | \boldsymbol{S}_t, \boldsymbol{A}_t\}]. \end{aligned} \tag{11}$$

The first line equals zero when $Q_i = Q_i^{\boldsymbol{\pi}}$ and $V_i = V_i(\boldsymbol{\pi})$. However, the second line depends on Q_i as well. As such, $(Q_i^{\boldsymbol{\pi}}, V_i(\boldsymbol{\pi}))$ might not necessarily be the minimizer of (11).

To resolve this issue, we consider first estimating the residual $\mathbb{E}\{R_{i,t} + Q_i(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}), \boldsymbol{S}_{t+1}) | \boldsymbol{S}_t, \boldsymbol{A}_t\} - V_i - Q_i(\boldsymbol{A}_t, \boldsymbol{S}_t)$ as a function of $\boldsymbol{A}_t$, $\boldsymbol{S}_t$, Q_i, V_i and then minimizing a regularized version of the squared residual. Under (7), we factorize $Q_i(\boldsymbol{a}, \boldsymbol{s})$ by $Q_i(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s}))$. This

yields the following optimization,

$$(12)\ (\widehat{V}_i(\boldsymbol{\pi}), \widehat{Q}_i) = \underset{(\eta, Q_i)\in\mathbb{R}\times\mathcal{Q}}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1} \widehat{g}_i^2(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t); \eta, Q_i) + \lambda\|Q_i\|_{\mathcal{Q}}^2,$$

where

$$(13)\quad \begin{aligned} \widehat{g}_i(\cdot,\cdot,\cdot,\cdot,\cdot;\eta,Q_i) = \underset{g\in\mathcal{G}}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1}\{R_{i,t} + Q_i(\pi_i(\boldsymbol{S}_{t+1}), \widetilde{S}_{i,t+1}) - \eta \\ -Q_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t)) - g(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))\}^2 + \mu\|g\|_{\mathcal{G}}^2, \end{aligned}$$

and μ and λ stand for some tuning parameters and $\|\cdot\|_{\mathcal{Q}}$, $\|\cdot\|_{\mathcal{G}}$ denote the corresponding RKHS norms. The purpose of adding the regularization terms in (12) and (13) is to prevent overfitting and guarantee the consistency of the estimator Q-function. Specifically, without the regularization term in (12), the estimated Q-function overfits to the noise and becomes inconsistent. The regularization term in (13) guarantees that the population limit of the RHS of (12) converges to the first line of (11). Without this regularization term, the coupled optimization problems reduces to (10), yielding a biased solution.

Next we derive the close-form expressions of $(\widehat{V}_i(\boldsymbol{\pi}), \widehat{Q}_i)$. Let $Z_{i,t} = (A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))^\top$ and $Z_{i,t}^* = (\pi_i(\boldsymbol{S}_{t+1}), \widetilde{S}_{i,t+1})^\top$. Let K_g and K_Q denote the reproducing kernels used to model g and Q, respectively. In our implementation, we use Gaussian RBF kernels to model these two functions. For a given Q_i and η, the optimizer of (11) $\widehat{g}_i$ can be represented by $\sum_{t=0}^{T-1}\widehat{\beta}_{i,t}K_g(Z_{i,t},\cdot)$. With some calculations, we obtain $\widehat{\boldsymbol{\beta}}_i = (\widehat{\beta}_{i,0}, \cdots, \widehat{\beta}_{i,T-1})^\top$ as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}}{\arg\min} \frac{1}{T}\sum_{t=0}^{T-1}\left\{R_{i,t} + Q_i(Z_{i,t}^*) - \eta - Q_i(Z_{i,t}) - \sum_{j=0}^{T-1}\beta_j K_g(Z_{i,j}, Z_{i,t})\right\}^2 + \mu\boldsymbol{\beta}^\top\boldsymbol{K}_g\boldsymbol{\beta} \\ = \frac{1}{T}\boldsymbol{\beta}^\top\{\boldsymbol{K}_g\boldsymbol{K}_g^\top + T\mu\boldsymbol{K}_g\}\boldsymbol{\beta} - \frac{2}{T}\boldsymbol{\beta}^\top\boldsymbol{K}_g(\boldsymbol{R} + \boldsymbol{Q}_i^* - \boldsymbol{Q}_i - \eta\mathbf{1}) \\ +\text{some terms that are independent of } \boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{K}_g = \{K_g(Z_{i,j_1}, Z_{i,j_2})\}_{j_1,j_2}$ and $\boldsymbol{R}$, $\boldsymbol{Q}_i^*$ and $\boldsymbol{Q}_i$ are the column vectors formed by elements in R_t, $Q_i(Z_{i,t}^*)$ and $Q_i(Z_{i,t})$, respectively. This allows us to derive a close-form expression for $\widehat{\boldsymbol{\beta}}_i$. See Appendix B for details. As a result, for a given Q_i and η, we have

$$\widehat{g}_i(Z_{i,t};\eta,Q_i) = \widehat{\boldsymbol{\beta}}_i^\top\boldsymbol{K}_g\boldsymbol{e}_t,$$

where $\boldsymbol{e}_t$ denotes the column vector with the t-th element equals to one and other elements equal to zero. As such,

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{g}_i^2(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t); \eta, Q_i) = \frac{1}{T}\widehat{\boldsymbol{\beta}}_i^\top\boldsymbol{K}_g\boldsymbol{K}_g^T\widehat{\boldsymbol{\beta}}_i.$$

Similarly, we can represent Q_i as $\sum_{t=0}^{2T-1}\widehat{\alpha}_{i,t}K_Q(\widetilde{Z}_{i,t},\cdot)$ where $\widetilde{Z}_{i,t}$ denotes the t-th element in the vector $(Z_{i,0}^\top, \cdots, Z_{i,T-1}^\top, Z_{i,0}^{*\top}, \cdots, Z_{i,T-1}^{*\top})^\top$. The closed-form expression of $\widehat{\boldsymbol{\alpha}}_i = (\widehat{\alpha}_{i,0}, \cdots, \widehat{\alpha}_{i,T-1})^\top$ can be similarly obtained. Details are given in Appendix B.

4. Simulations. In this section, we design an example to simulate the environment of a ride-sharing platform. Specifically, we consider orders and drivers operating in a map of 5×5 spatial grids. For each grid, we design three time-varying variables to construct the state. Let $D_{i,t}$ and $O_{i,t}$, respectively, denote the number of drivers and orders in the i-th

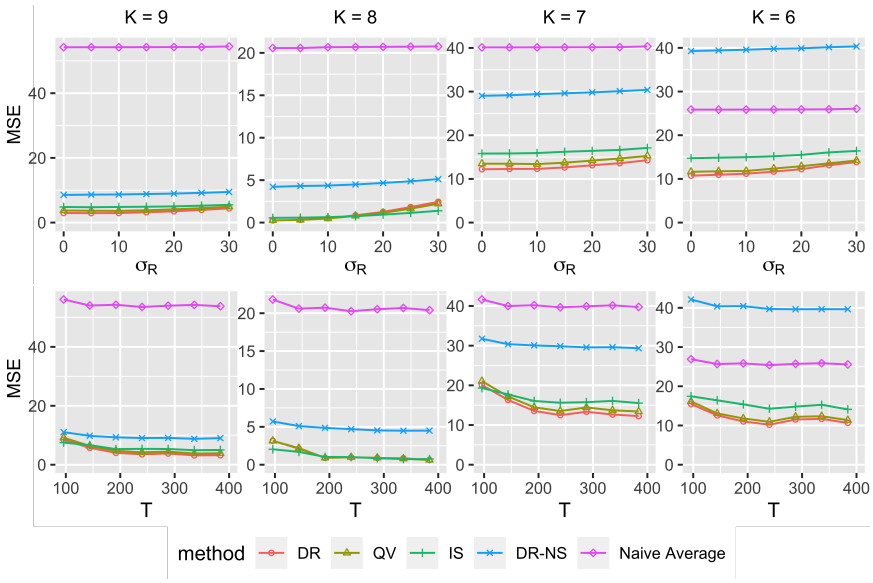

Fig 4: Mean squared errors of different value estimates, aggregated over 100 simulations. T is set to 336 (the experiment lasts for two weeks and each hour is treated as one time unit) in the top plots and σ_R is set to 15 in the bottom plots. MSEs of DR-NM are larger than 1000 for all choices of σ_R and T and thus are not plotted.

grid during the time interval $(t-1, t]$. In practice, these two factors are known to have large impact on the driver income and customer satisfaction. The last variable $M_{i,t}$ measures the degree of mismatch between orders and drivers in the i-th grid at time t. Specifically, we set $M_{i,t} = 0.5\{1 - |D_{i,t} - O_{i,t}|/(1 + D_{i,t} + O_{i,t})\} + 0.5M_{i,t-1}$. Given the state vector at time $t+1$, we generate the reward $R_{i,t}$ from the following model:

$$R_{i,t} = M_{i,t+1} \min(D_{i,t+1}, O_{i,t+1}) + \varepsilon_{i,t}, \tag{14}$$

where $\{\varepsilon_{i,t}\}_{i,t}$ are i.i.d. $\mathcal{N}(0, \sigma_R^2)$. At each time t, the platform can decide whether to implement a certain driver-side subsidizing policy to the ith spatial unit or not. This yields a binary action $A_{i,t}$. In our experiment, $\{A_{i,t}\}_{i,t}$ are i.i.d. according to a Bernoulli distribution with success probability 0.5.

Orders are simulated in the following manner. For $i = 1, \ldots, 25$, we first randomly generate μ_i from $N(100, 25^2)$. Then we independently generate $O_{i,t}$'s from a Poisson distribution with expectation μ_i. Thus, each μ_i represents the average number of orders during each time unit in region i. We plot these μ_i's in Figure 6 (a).

Drivers are simulated in the following manner. Initially, we put 80 drivers in each grid. At each time, drivers will be attracted to nearby regions that either implement the subsidizing policy or has large number of orders. To characterize this effect, at each time, we assign an attraction rate parameter to each region as $\nu_{i,t} = 1.5\exp(A_{i,t}) + 0.5(O_{i,t}/D_{i,t})$. At time $t+1$, drivers are more likely to move to neighborhood regions with large attraction rates. Specifically, we set $D_{i,t+1} = \nu_{i,t} D_{i,t} (\sum_{j \in \mathcal{N}(i)} \nu_{j,t})^{-1}$.

Under the current setup, to increase the long-term reward, we can implement subsidizing policies in regions with large μ_i's. Specifically, we focus on four nondynamic policies $\{\boldsymbol{\pi}_K\}_{K \in \{6,7,8,9\}}$. Under $\boldsymbol{\pi}_K$, the subsidizing policies will be implemented in the top K regions with largest μ_i's at each time. We are interested in evaluating the values $V(\boldsymbol{\pi}_K)$ for $K = 6, 7, 8, 9$. The true value of $V(\boldsymbol{\pi}_K)$ can be approximated via Monte Carlo simulations. Specifically, in each simulation, we generate data trajectories $\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1 \le i \le N, 0 \le t < M}$

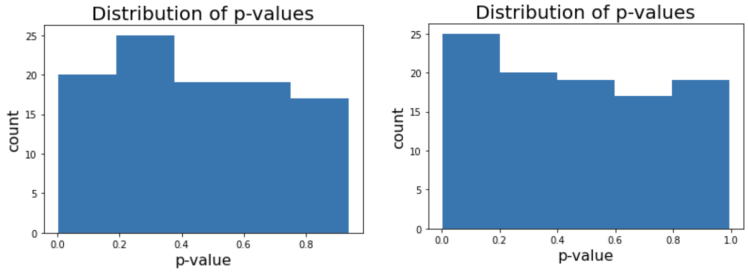

Fig 5: Distribution of p-values for testing the mean-field approximation assumption associated with two randomly selected regions.

under the target policy $\boldsymbol{\pi}_K$ for some sufficiently large integer $M > 0$, and calculate the average reward $(NM)^{-1}\sum_{i=1}^{N}\sum_{t=0}^{M-1} R_{i,t}$. Then we aggregate these rewards over hundreds of simulations and treat it as the true value of $V(\boldsymbol{\pi})$. These estimated true values ranges from 55 to 60, with standard errors ranges from 0.05 to 0.35, across different settings. Comparisons are made among the following methods:
(a) The doubly-robust estimator (denoted by DR) proposed in Section 3;
(b) The IS based estimator (denoted by IS) proposed in Section 3;
(c) The doubly-robust estimator without considering the spatial dependence (denoted by DR-NS);
(d) The doubly-robust estimator in (6) without the mean field approximation (denoted by DR-NM);
(e) The Q-function based value estimator $N^{-1}\sum_{i=1}^{N}\widehat{V}_i(\boldsymbol{\pi})$ where each $\widehat{V}_i(\boldsymbol{\pi})$ is computed according to Section 3.4 (denoted by QV);
(f) A naive average of all immediate rewards.

To implement the proposed estimator, we set the mean-field functions to the averaged state and action over the neighbors, as in (3). Before presenting the results, we first investigate the validity of the mean-field approximation assumption with such a choice of mean-field functions. In our simulations, the reward is generated according to the additive noise model (see (14)). As we have commented, it suffices to test the conditional independence of $(\pi_i(\boldsymbol{S}_{t+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{t+1})), \widetilde{S}_{i,t+1})$ and $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. Notice that according to the definition of $\boldsymbol{\pi}_K$, the evaluation policies depend only on μ_i's (which are fixed over different simulations) and are thus state-agnostic, e.g., $\boldsymbol{\pi}_K(\boldsymbol{S}_t) = \boldsymbol{\pi}_K$. As such, $\pi_i(\boldsymbol{S}_{t+1})$ and $m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{t+1}))$ are independent of $\boldsymbol{A}_t$ and $\boldsymbol{S}_t$. It suffices to test the conditional independence of $\widetilde{S}_{i,t+1}$ and $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. Here, we adopt the forward-backward learning procedure developed by Shi et al. (2020b). There test does not impose parametric model assumptions on the transition dynamic and is consistent even in high-dimensional settings. In Figure 5, we report the distributions of p-values for testing the mean-field approximation assumption associated with two randomly selected regions. It can be seen that the p-values are approximately uniformly distributed. This implies that the mean-field approximation assumption is likely to hold in simulations.

Our experiments were run on an c5d.9xlarge instance on the AWS EC2 platform, with 36 cores and 72GB RAM. Computing our estimator for one target policy takes roughly one minute. To calculate DR-NS, for each spatial unit, we separately calculate a doubly-robust estimator based on its own state-action-reward triplets. The final estimator is defined as a simple average of all these estimators. Mean square errors (MSEs) of value estimators in (a)-(f) are reported in Figure 4 (b) with different choices of σ_R and T. MSEs of DR-NM are larger than 1000 for all choices of σ_R and T and thus are not plotted.

We summarize our findings below: (i) The proposed IS and DR estimators achieve smaller MSEs than DR-NS, DR-NM and Naive Average in all cases. In addition, we find that the standard errors of these MSEs are much smaller than the differences between MSEs. Consequently, our methods perform significantly better than the competing baselines. As commented before, this is due to the fact that DR-NS and the naive average estimator ignore interactions among different agents while DR-NM suffers from the curse of dimensionality; (ii) In general, MSEs of our proposed estimators increase with σ_R and decreases with T, though the trend is not that obvious; (iii) DR outperforms IS and QV in most cases. Although the MSEs of DR and QV are very similar in Figure 4, we conduct paired two-sample t-test to test whether the MSE of DR is strictly smaller than that of QV, and find that the tests are significant in most cases. See Appendix E of the supplementary article for details.

5. Applications. We apply the proposed method to a real dataset from a ride-sharing company, to examine the effectiveness of applying driver-side subsidizing policies and customer recommendation programs. The data is from a randomized experiment, conducted at a given city, partitioned into eight irregular spatial regions, as shown in Figure 1. Thirty-minutes is defined as one time unit. Figure 6 (b) depicts the number of orders $O_{i,t}$ within each spatiotemporal unit, averaged across days. The total Gross Merchandise Volume (GMV) within each spatiotemporal unit is set to the immediate reward $R_{i,t}$.

Due to data confidentiality, we are not able to use the raw data. Here, we simulate $D_{i,t}$, $O_{i,t}$, $R_{i,t}$ and the degree of mismatch between orders and drivers $M_{i,t}$ for $i = 1, \ldots, 8$ and $t = 0, \ldots, T-1$. Specifically, we conduct two real data based simulations to evaluate two promotion strategies, one offered to the drivers and another to the passengers. As commented in the introduction, driver-side subsidies will change the spatial distribution of drivers, inducing interference effects in both space and time. Applying recommendation programs to certain passengers will increase their chances of requesting orders in the future, inducing interference effects in time.

In the first simulation experiment, we aim to evaluate driver-side subsidizing policy. The numbers of orders $\{O_{i,t}\}_{i,t}$ are independently generated. The spatial distribution of orders are identical across days and the expected number of orders is set to the historical average. The numbers of drivers $\{D_{i,t}\}_{i,t}$ are affected by both the subsidizing action and the spatial distribution of orders. We use a similar generative model as in Section 4 with the model parameters estimated using the real dataset. In the second simulation experiment, we aim to evaluate the causal effect of applying passenger recommendation programs. The distribution of $\{O_{i,t}\}_{i,t}$ is set to depend on the action. The distribution of $\{D_{i,t}\}_{i,t}$, however, is set to depend on the orders only. For both experiments, the data generative models are chosen such that the distribution of these generative variables are very similar to that of the observed data.

Similar to the simulation study, we first apply the forward-backward learning procedure to test the mean-field approximation assumption. Recall that we partition the city into eight spatial regions. The empirical rejection probabilities associated with these regions are 0.03, 0.09, 0.06, 0.07, 0.04, 0.08, 0.03 and 0.1, respectively, under the significance level 0.05. As such, we expect this assumption holds in our data application as well.

It can be seen from Figure 6 that the data are time-varying within each day. To guarantee the transition matrix is homogeneous in time (see MA), we define the state $S_{i,t} = (D_{i,t}, O_{i,t}, M_{i,t}, T_t)^\top$ for any i and t, where the variable T_t denotes the time of the day. Since one time unit consists of thirty-minutes, T_t satisfies $T_t = T_{t+48}$ for any t. The Markov chain will converge to its limiting distribution under mild conditions as the number of days approaches infinity (Lloyd, 1977). As such, the proposed method remains valid. We are interested in evaluating the values under the driver-side policies (i) - (iii) in the first simulation and the passenger-side policies (iv) - (vi) in the second simulation.

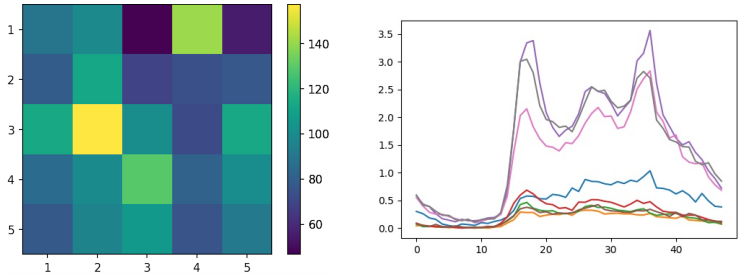

Fig 6: (a) visualization of μ_i; (b) number of orders within each spatiotemporal unit, averaged across days. The numbers in (b) are scaled due to sensitivity and privacy concerns.

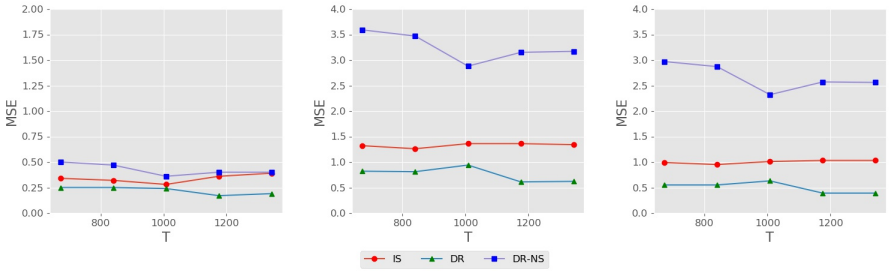

Fig 7: Mean squared errors of different value estimates, aggregated over 100 simulations for driver-side subsidizing policies. The target policies are spatially adaptive, timely adaptive and sptiotemporally adaptive, from left to right. MSEs of DR-NM are over 100 in all cases and are hence not reported.

(i) Driver-side spatially adaptive policy: the four regions that have the largest numbers of orders on average will receive the subsidizing policy all the time.
(ii) Driver-side timely adaptive policy: all regions during peak hours will receive the subsidizing policy.
(iii) Driver-side spatiotemporally adaptive policy: the four regions that have the largest number of orders on average will receive the subsidizing policy during peak hours.
(iv) Passenger-side spatially adaptive policy: the four regions that have the most vacant drivers will receive the subsidizing policy all the time.
(v) Passenger-side timely adaptive policy: all regions during time periods with the most vacant drivers will receive the subsidizing policy.
(vi) Passenger-side spatiotemporally adaptive policy: the four regions that have the most vacant drivers on average will receive the subsidizing policy during non-peak hours.

For each of the target policy, we compute the MSEs of value estimators constructed based on (a)-(d), with different choices of T. The naive average (e) is not considered since the estimator is meaningless when the target differs from the behavior policy. MSEs of DR-NM are over 100 in all cases and are hence not reported. MSEs of other methods are plotted in Figure 7 for the first simulation and Figure 8 for the second simulation. It can be seen that DR outperforms other competitors in most cases. It is worth mentioning that DR-NS performs better than IS when evaluating (v) and (vi). This is partly because applying passenger-side recommendation program to a given region only affect the number of orders within that region. The interference effect in space is moderate in this case.

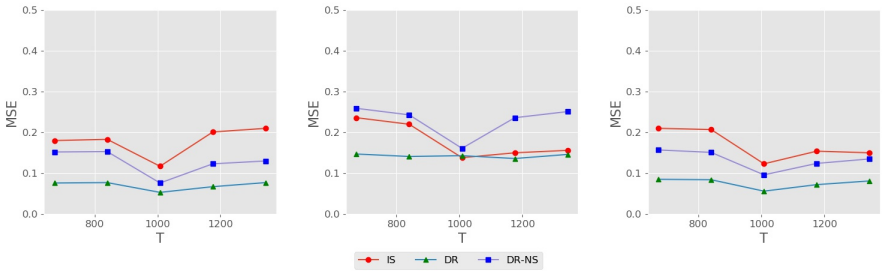

Fig 8: Mean squared errors of different value estimates, aggregated over 100 simulations for passenger-side subsidizing policies. The target policies are spatially adaptive, timely adaptive and sptiotemporally adaptive, from left to right.

6. Discussion. In this article, we introduce a MARL framework for spatiotemporal treatment effects evaluation and develop a novel off-policy value estimator in MARL. The validity of our method relies on SRA. It would be practically interesting to extend our proposal to settings where SRA is violated. We leave it for future research.

For simplicity, we set the mean-field function to be the averaged state or action over the neighbors in our numerical experiments. Such a choice works well in our experiments as the MSE of the proposed estimator is much smaller than those of the estimators without considering the spatial dependence or the mean-field approximation. In addition, we find that the resulting mean-field approximation assumption is likely to hold in our numerical experiments. However, it remains unclear whether such a choice of the mean-field function is optimal. We next propose two methods for adaptively selecting the mean-field function in practice. The first method requires a canddiate set of mean-field functions. Given each mean-field function in the candidate set, we can apply existing state-of-the-art conditional independence test to compute its p-value. We next select the mean-field function with the largest p-value. The second method parametrize the mean-field function using some universal approximator (e.g., neural networks) and estimate the associated parameters by minimizing some empirical measures of the conditional dependence of $(\pi_i(\boldsymbol{S}_{t+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{t+1})), S_{i,t+1}, m_i^s(\boldsymbol{S}_{t+1}))$ and $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$, such as the maximum mean discrepancy (see e.g., Fukumizu et al., 2007). Since the estimated mean-field function minimizes these measures, it is likely to satisfy the conditional independence assumption. It is practically interesting to further investigate these methods. However, it is beyond the scope of the current paper. We leave it for future research.

Let $\mathbf{0}$ and $\mathbf{1}$ denote two non-dynamic policies that assign Treatments 0 and 1 to each region at any time. We can decompose the value difference $V(\mathbf{1}) - V(\mathbf{0})$ as the sum of direct effects (DE) and indirect effects (IE). Specifically, it follows from CMIA that

$$V(\mathbf{1}) - V(\mathbf{0}) = \lim_{t\to\infty} \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t} \mathbb{E}\{R_{i,j}^*(\mathbf{1}) - R_{i,j}^*(\mathbf{0})\}$$

$$= \underbrace{\lim_{t\to\infty} \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t} \mathbb{E}\{r_i(\mathbf{1}, \boldsymbol{S}_j^*(\mathbf{0})) - r_i(\mathbf{0}, \boldsymbol{S}_j^*(\mathbf{0}))\}}_{\text{DE}}$$

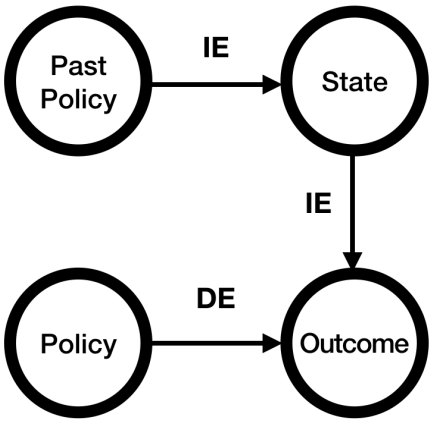

Fig 9: Graphical illustration of direct and indirect effects.

$$\underbrace{+\lim_{t\to\infty}\frac{1}{Nt}\sum_{i=1}^{N}\sum_{j=0}^{t}\mathbb{E}\{r_i(\mathbf{1},\boldsymbol{S}_j^*(\mathbf{1}))-r_i(\mathbf{1},\boldsymbol{S}_j^*(\mathbf{0}))\}}_{\text{IE}},$$

where $\boldsymbol{S}_t^*(\mathbf{0})$ and $\boldsymbol{S}_t^*(\mathbf{1})$ denote the potential states under Policies $\mathbf{0}$ and $\mathbf{1}$, respectively. See Figure 9 for a graphical illustration. The DE represents the sum of the short-term treatment effects on the immediate outcome over time assuming that the baseline policy is being employed in the past. In contrast, IE characterizes the carryover effects of past policies that work through the state vector. It is practically interesting to extend the current proposal to estimating DE and IE. However, this is beyond the scope of the current paper. We leave it for future research.

REFERENCES

Armstrong, M. (2006). Competition in two-sided markets. *The RAND journal of economics*, 37(3):668–691.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc.*, 113(521):230–240.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*.

Bhattacharya, R., Malinsky, D., and Shpitser, I. (2019). 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.

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

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.*, 113(523):1112–1121.

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

Cai, H., Shi, C., Song, R., and Lu, W. (2021). Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300.

Chakraborty, B., Laber, E. B., and Zhao, Y.-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.*, 19(3):317–343.

Chen, E. Y., Hu, Z. T., Song, R., and Jordan, M. I. (2020). Heterogeneous reinforcement learning with offline data: Estimation and inference.

Chen, X. and Qi, Z. (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. *arXiv preprint arXiv:2201.06169*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Dempsey, W., Liao, P., Kumar, S., and Murphy, S. A. (2017). The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *arXiv preprint arXiv:1711.03587*.

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

Ertefaie, A. (2014). Constructing dynamic treatment regimes in infinite-horizon settings. *arXiv preprint arXiv:1406.0764*.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

Fang, E. X., Wang, Z., and Wang, L. (2021). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, accepted.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.*, 17:Paper No. 139, 66.

Frenken, K. and Schor, J. (2017). Putting the sharing economy into perspective. *Environmental Innovation and Societal Transitions*, 23:3–10.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496.

Hagiu, A. and Wright, J. (2019). The status of workers and platforms in the sharing economy. *Journal of Economics & Management Strategy*, 28:97–108.

Halloran, M. E. and Hudgens, M. G. (2016). Dependent happenings: a recent methodological review. *Current epidemiology reports*, 3(4):297–305.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2019). Personalized policy learning using longitudinal mobile health data. *Journal of the American Statistical Association*, accepted.

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

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

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

Jin, S. T., Kong, H., Wu, R., and Sui, D. Z. (2018). Ridesourcing, the sharing economy, and the future of cities. *Cities*, 76:96–104.

Kallus, N. and Uehara, M. (2019). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

Laber, E. B., Meyer, N. J., Reich, B. J., Pacifici, K., Collazo, J. A., and Drake, J. M. (2018). 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.

Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G., and Yang, Q. (2011). Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 63–68. IEEE.

Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. (2022a). Reinforcement learning in possibly nonstationary environments. *arXiv preprint arXiv:2203.01707*.

Li, Y., Wang, C.-h., Cheng, G., and Sun, W. W. (2022b). Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*.

Liao, P., Klasnja, P., and Murphy, S. (2020). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, accepted.

Liao, P., Qi, Z., and Murphy, S. (2022). Batch policy learning in average reward Markov decision processes. *Annals of Statistics*, accepted.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366.

Lloyd, E. H. (1977). Reservoirs with seasonally varying Markovian inflows and their first passage times.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2019). Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, accepted.

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

Matsouaka, R. A., Li, J., and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics*, 70(3):489–499.

Meng, H., Zhao, Y.-Q., Fu, H., and Qiao, X. (2020). Near-optimal individualized treatment recommendations. *arXiv preprint arXiv:2004.02772*.

Miao, F., Han, S., Lin, S., Stankovic, J. A., Zhang, D., Munir, S., Huang, H., He, T., and Pappas, G. J. (2016). Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach. *IEEE Transactions on Automation Science and Engineering*, 13(2):463–478.

Mo, W., Qi, Z., and Liu, Y. (2020). Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, pages 1–16.

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

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*.

Ning, B., Ghosal, S., Thomas, J., et al. (2019). Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis*, 14(1):1–28.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

Reich, B. J., Yang, S., Guan, Y., Giffin, A. B., Miller, M. J., and Rappold, A. (2020). A review of spatial causal inference methods for environmental and epidemiological applications. arXiv:2007.02714v1.

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

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

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

Rysman, M. (2009). The economics of two-sided markets. *Journal of Economic Perspective*, 23:125–143.

Shi, C., Fan, A., Song, R., and Lu, W. (2018a). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925.

Shi, C., Lu, W., and Song, R. (2020a). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67.

Shi, C., Song, R., Lu, W., and Fu, B. (2018b). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 80(4):681.

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

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. (2020b). Does the Markov decision process fit the data: testing for the Markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR.

Shi, C., Wang, X., Luo, S., Zhu, H., Ye, J., and Song, R. (2022a). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, pages 1–13.

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

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015). Penalized Q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

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

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

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497.

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

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228–1242.

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

Wang, L., Yang, Z., and Wang, Z. (2020). Provably efficient causal reinforcement learning with confounded observational data. *arXiv preprint arXiv:2006.12311*.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254.

Wu, Y. and Wang, L. (2020). Resampling-based confidence intervals for model-free robust inference on optimal treatment regimes. *Biometrics*.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018). Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang., A. (2020). A survey on causal inference. arXiv:2002.02770.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018.

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

Zhang, D., Sun, L., Li, B., Chen, C., Pan, G., Li, S., and Wu, Z. (2014). Understanding taxi service strategies from taxi gps traces. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):123–135.

Zhang, K., Yang, Z., and Başar, T. (2019). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*.

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018). Estimation of optimal treatment regimes using lists. *J. Amer. Statist. Assoc.*, 113(524):1541–1549.

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.*, 107(499):1106–1118.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.*, 110(510):583–598.

Zhou, W., Zhu, R., and Qu, A. (2021). Estimating optimal infinite horizon dynamic treatment regimes via pt-learning. *arXiv preprint arXiv:2110.10719*.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675.

APPENDIX A: THEORETICAL RESULTS

This section is organized as follows. We first introduce a key lemma that establishes an exponential inequality for the suprema of empirical processes under weak dependence. We next discuss the statistical properties of the proposed estimator.

A.1. A key lemma. We briefly introduce the setup before presenting the lemma. Let $\{Z_t : t \geq 0\}$ be a stationary β-mixing process whose β-mixing coefficients are given by $\{\beta(q) : q \geq 0\}$. Let $\mathcal{F}$ be a pointwise measurable class of functions that take Z_t as input with a measurable envelope function F. For any $f \in \mathcal{F}$, suppose $\mathbb{E}f(Z_0) = 0$. Let $\sigma^2 > 0$ be a positive constant such that $\sup_{f\in\mathcal{F}} \mathbb{E}f^2(Z_0) \leq \sigma^2 \leq \mathbb{E}F^2(Z_0)$. We focus providing an exponential inequality for the empirical process $\sup_{f\in\mathcal{F}} |\sum_{t=0}^{T-1} f(Z_t)|$.

Toward that end, we introduce the notion of the VC type class (Chernozhukov et al., 2014, Definition 2.1). For any probability measure Q, let e_Q denote a semi-metric on $\mathcal{F}$ such that $e_Q(f_1, f_2) = \|f_1 - f_2\|_{Q,2} = \sqrt{Q|f_1 - f_2|^2}$. An ϵ-net of the space $(\mathcal{F}, e_Q)$ is a subset $\mathcal{F}_\epsilon$ of $\mathcal{F}$, such that for every $f \in \mathcal{F}$, there exists some $f_\epsilon \in \mathcal{F}_\epsilon$ satisfying $e_Q(f, f_\epsilon) < \epsilon$. We say

that $\mathcal{F}$ is a VC type class with envelope F, if there exist constants $c_0 > 0, c_1 \geq 1$, such that $\sup_Q N(\mathcal{F}, e_Q, \epsilon\|F\|_{Q,2}) \leq (c_0/\epsilon)^{c_1}$, for all $0 < \epsilon \leq 1$, where the supremum is taken over all finitely discrete probability measures on the support of $\mathcal{F}$, and $\mathbb{N}(\mathcal{F}, e_Q, \epsilon\|F\|_{Q,2})$ is the infimum of the cardinality of $\epsilon\|F\|_{Q,2}$-nets of $\mathcal{F}$. We refer to c_1 as the VC index of $\mathcal{F}$.

LEMMA 3. *Suppose the envelop function is uniformly bounded by some constant $M > 0$. In addition, suppose $\mathcal{F}$ belongs to the class of VC-type class such that $\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leq (A/\varepsilon)^\nu$ for some $A \geq e, \nu \geq 1$. Then there exist some constants $c, C > 0$ such that*

$$Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > c\sqrt{\nu q\sigma^2 T\log\left(\frac{AM}{\sigma}\right)} + c\nu M\log\left(\frac{AM}{\sigma}\right) + cq\tau + Mq\right)$$
$$\leq Cq\exp\left(-\frac{\tau^2 q}{CT\sigma^2}\right) + Cq\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},$$

for any $\tau > 0$, $1 \leq q < T/2$.

As commented in the introduction, Lemma 3 is useful for finite-sample analysis of machine learning estimates under weak dependence. It allows us to derive a sharp bound on the difference between the proposed value estimator and the oracle estimator.

A.2. Statistical performance guarantees. To derive these theoretical results, we need one additional condition to characterize the dependence between observations over time.

(A1) The process $\{(\boldsymbol{S}_t, \boldsymbol{A}_t)\}_{t\geq 0}$ is strictly stationary. Its β-mixing coefficients $\{\beta(q)\}_{q\geq 0}$ (see e.g., Bradley, 2005, for definition) satisfy $\beta(q) \leq \kappa_0\rho^q$ for some constants $\kappa_0 > 0$ and $0 < \rho < 1$.

When (A1) holds and the initial distribution of $\{\boldsymbol{S}_t\}_{t\geq 0}$ equals to its stationary distribution, the stationarity condition in (A1) is automatically satisfied. The second part of (A1) holds when $\{(\boldsymbol{S}_t, \boldsymbol{A}_t)\}_{t\geq 0}$ satisfies geometric ergodicity. Geometric ergodicity is weaker than the uniform ergodicity condition imposed in the existing reinforcement learning literature (Bhandari et al., 2018; Zou et al., 2019).

In addition, let $V_i^*(\boldsymbol{\pi})$ and ω_i^* be the population limit of $\widehat{V}_i(\boldsymbol{\pi})$ and $\widehat{\omega}_i$, respectively. We require $V_i^*(\boldsymbol{\pi}) = V_i(\boldsymbol{\pi})$ when (7) holds and $\omega_i^* = \omega_i$ when (4) holds.

THEOREM 1. *Suppose (A1) hold, $NTVar\{\widehat{V}^{DR*}(\boldsymbol{\pi})\} \to \sigma^2 > 0$ and $T \to \infty$. Suppose that $\{R_{i,t}, \bar{Q}_i, \omega_i^*, V_i(\boldsymbol{\pi}) : 1 \leq i \leq N, t \geq 0\}$ are uniformly bounded from infinity and the set of functions $\{b_i : 1 \leq i \leq N\}$ are uniformly bounded from zero. Then as either* (4) *or* (7) *holds, $\sqrt{NT}\{\widehat{V}^{DR*}(\boldsymbol{\pi}) - V(\boldsymbol{\pi})\} \xrightarrow{d} N(0, \sigma^2)$, where $\xrightarrow{d}$ denotes converge in distribution.*

We next investigate the statistical properties of our estimator $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$. Our theories are generally applicable to a class of density ratio and Q-function estimators that satisfy certain mild conditions. We summarize these conditions in (A2) and (A3) and present them below.

(A2)(i) $\sum_{i=1}^N |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})|/N = o_p(1)$; (ii) $\widehat{Q}_{i,\boldsymbol{\pi}} \in \mathcal{Q}$, $\widehat{\omega}_i \in \mathcal{W}$ almost surely for any i where the function classes $\mathcal{Q}$ and $\mathcal{W}$ belong to VC type classes with VC indices bounded by $\nu = O(T^\kappa)$ for some $0 \leq \kappa < 1$, and envelope functions are bounded by some constant M. (iii) $\max_{i,a}\int_{\tilde{s}_i} |\widehat{Q}_i(a, \tilde{s}_i) - Q_i^*(a, \tilde{s}_i)|^2 p_{i,b}(\tilde{s}_i)d\tilde{s}_i = o_p(1)$, $\max_i \int_{\tilde{s}_i} |\widehat{\omega}_i(\tilde{s}_i) - \omega_i^*(\tilde{s}_i)|^2 p_{i,b}(\tilde{s}_i)d\tilde{s}_i = o_p(1)$.

(A3)(i) $\max_i |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})|^2 = o_p((NT)^{-1/2})$; (ii) $\max_{i,a} \int_{\tilde{s}_i} |\widehat{Q}_i(\tilde{s}_i) - Q_i^*(a, \tilde{s}_i)|^2 p_{i,b}(\tilde{s}_i) d\tilde{s}_i = o_p((NT)^{-1/2})$; (iii) $\max_i \int_{\tilde{s}_i} |\widehat{\omega}_i(\tilde{s}_i) - \omega_i^*(\tilde{s}_i)|^2 p_{i,b}(\tilde{s}_i) d\tilde{s}_i = o_p((NT)^{-1/2})$; (iv) $T \gg N\nu^2 \log^4(NT)$.

Condition (A2) requires $\widehat{V}_i(\boldsymbol{\pi})$, $\widehat{Q}_i$ and $\widehat{\omega}_i$ to be consistent whereas (A3) requires these estimators to converge at a certain rate. When using RKHS or neural networks to estimate the Q-function and the density ratio, the corresponding convergence rates have been established (Fan et al., 2020; Kallus and Uehara, 2019; Liao et al., 2022). The conditions on $\mathcal{Q}$ and $\mathcal{W}$ in (A2) are mild as these function classes are user-specified.

THEOREM 2 (doubly-robustness). *Suppose the conditions in Theorem 1 hold. Suppose (A2) holds. Then as either* (4) *or* (7) *holds, we have* $\widehat{V}^{DR}(\boldsymbol{\pi}) - V(\boldsymbol{\pi}) = o_p(1)$.

THEOREM 3 (oracle property). *Suppose the conditions in Theorem 2 hold. Suppose (A3) holds. Then when both* (4) *and* (7) *hold, we have* $\sqrt{NT}\{\widehat{V}^{DR}(\boldsymbol{\pi}) - V(\boldsymbol{\pi})\} \xrightarrow{d} N(0, \sigma^2)$.

Theorem 3 implies that the asymptotic variance of the doubly-robust estimator is the same as that of the oracle estimator, when both mean-field approximations are valid. The explicit formula for σ^2 is given in Appendix D.2 of the supplementary article. We also remark that different from the i.i.d. case considered in classical semiparametric statistics, there is no guarantee that the asymptotic variance of the doubly-robust estimator will be smaller than or equal to that of the importance sampling estimator under our setting where observations are time dependent.

APPENDIX B: MORE ON THE LEARNING PROCEDURE

We first present a close-form expression for $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\boldsymbol{\alpha}}_i$. We next discuss the case where the behavior policy is unknown.

B.1. Closed-form for $\widehat{\boldsymbol{\beta}}_i$. Notice that $\boldsymbol{K}_g$ is symmetric, by some calculations, we obtain

$$\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{K}_g \boldsymbol{K}_g^\top + T\mu \boldsymbol{K}_g)^{-1} \boldsymbol{K}_g (\boldsymbol{R} + \boldsymbol{Q}_i^* - \boldsymbol{Q}_i - \eta \mathbf{1}) = (\boldsymbol{K}_g + T\mu \boldsymbol{I})^{-1} (\boldsymbol{R} + \boldsymbol{Q}_i^* - \boldsymbol{Q}_i - \eta \mathbf{1}).$$

B.2. Closed-form for $\widehat{\boldsymbol{\alpha}}_i$. Let $\boldsymbol{K}_Q$ denotes a $2T \times 2T$ matrix where its (j_1+1, j_2+1)th element is given by $K_Q(\widetilde{Z}_{i,j_1}, \widetilde{Z}_{i,j_2})$, we have

$$Q_i(Z_{i,t}) = \boldsymbol{\alpha}_i^\top \boldsymbol{K}_Q \boldsymbol{e}_t \ \text{ and } \ Q_i(Z_{i,t}^*) = \widehat{\boldsymbol{\alpha}}_i^\top \boldsymbol{K}_Q \boldsymbol{e}_{t+T}.$$

It follow that

$$\boldsymbol{Q}_i^* - \boldsymbol{Q}_i = \underbrace{[-\boldsymbol{I}_T, \boldsymbol{I}_T]}_{\boldsymbol{C}} \boldsymbol{K}_Q \widehat{\boldsymbol{\alpha}}_i.$$

Note that $\boldsymbol{K}_Q$ is symmetric. Let $\boldsymbol{E} = \boldsymbol{K}_g^\top \{\boldsymbol{K}_g + (T-1)\mu \boldsymbol{I}\}^{-1}$, $\widehat{\boldsymbol{\alpha}}_i$ corresponds to the solution of the following optimization problem,

$$\widehat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \boldsymbol{\alpha} - \eta \mathbf{1})^\top \boldsymbol{E}^\top \boldsymbol{E} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \boldsymbol{\alpha} - \eta \mathbf{1}) + T\lambda \boldsymbol{\alpha}^\top \boldsymbol{K}_Q \boldsymbol{\alpha}.$$

Taking derivatives with respect to $\boldsymbol{\alpha}$ and η, we obtain

$$(\widehat{\boldsymbol{\alpha}}_i, \widehat{V}_i(\boldsymbol{\pi}))^\top = -([\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}]^\top \boldsymbol{E}^\top \boldsymbol{E} [\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}] + [T\lambda \boldsymbol{K}_Q, \mathbf{0}; \mathbf{0}^\top, 0])^{-1} [\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}] \boldsymbol{E}^\top \boldsymbol{E} \boldsymbol{R}.$$

B.3. Unknown behavior policy. Note that $b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t}) = \mathbb{E}\{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))|\widetilde{S}_{i,t}\}$. It can thus be learned by applying machine learning algorithms to datasets with responses $\{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t))) : 0 \le t < T\}$ and predictors $\{\widetilde{S}_{i,t} : 0 \le t < T\}$.

APPENDIX C: MORE ON THE MEAN-FIELD APPROXIMATION

We first show that (7) holds when (4) is satisfied and that $(\pi_i(\boldsymbol{S}_{t+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{t+1})), \widetilde{S}_{i,t+1})$ is conditionally independent of $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ given $(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$. When (4) is satisfied, it follows that

$$Q_i^{\boldsymbol{\pi}}(\boldsymbol{a}, \boldsymbol{s}) = \sum_{t \ge 0} \gamma^t \mathbb{E}^{\boldsymbol{\pi}}\{\bar{r}_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\}.$$

It suffices to show that the conditional mean function $\mathbb{E}^{\boldsymbol{\pi}}\{\bar{r}_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\}$ depends on $(\boldsymbol{a}, \boldsymbol{s})$ only through $(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s}))$, for any $t \ge 0$. When $t = 0$, this assertion is automatically satisfied by (4). When $t > 0$, we have

$$\mathbb{E}^{\boldsymbol{\pi}}\{\bar{r}_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\}.$$

It suffices to show that for any square integrable function h, $\mathbb{E}^{\pi}\{h(\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)), \widetilde{S}_{i,t})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\}$ depends on $(\boldsymbol{a}, \boldsymbol{s})$ only through $(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s}))$, for any $t \ge 1$.

We next prove this assertion by induction. When $t = 1$, this assertion is automatically satisfied by condition. Suppose the assertion holds with $t \le j$. It suffices to show it is satisfied for $t = j + 1$. Under MA and the given conditions, for any square integrable function h, we have

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\pi}}\{h(\pi_i(\boldsymbol{S}_{j+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{j+1})), \widetilde{S}_{i,j+1})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\} \\
&= \mathbb{E}^{\boldsymbol{\pi}}[\mathbb{E}^{\boldsymbol{\pi}}\{h(\pi_i(\boldsymbol{S}_{j+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{j+1})), \widetilde{S}_{i,j+1})|\boldsymbol{A}_j, \boldsymbol{S}_j\}|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}] \\
&= \mathbb{E}^{\boldsymbol{\pi}}[\mathbb{E}^{\boldsymbol{\pi}}\{h(\pi_i(\boldsymbol{S}_{j+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{j+1})), \widetilde{S}_{i,j+1})|A_{i,j}, m_i^a(\boldsymbol{A}_j), \widetilde{S}_{i,j}\}|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}] \\
&= \mathbb{E}^{\boldsymbol{\pi}}[\mathbb{E}^{\boldsymbol{\pi}}\{h(\pi_i(\boldsymbol{S}_{j+1}), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_{j+1})), \widetilde{S}_{i,j+1})|\pi_i(\boldsymbol{S}_j), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_j)), \widetilde{S}_{i,j}\}|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}].
\end{aligned}$$

The last line is a function of $(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s}))$ only, given that $\mathbb{E}^{\boldsymbol{\pi}}\{h(\pi_i(\boldsymbol{S}_j), m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_j)), \widetilde{S}_{i,j})|\boldsymbol{A}_0 = \boldsymbol{a}, \boldsymbol{S}_0 = \boldsymbol{s}\}$ depends on $(\boldsymbol{a}, \boldsymbol{s})$ only through $(a_i, m_i^a(\boldsymbol{a}), s_i, m_i^s(\boldsymbol{s}))$. This completes the proof.

APPENDIX D: PROOFS

We use c and C to denote some generic constants whose values are allowed to vary from place to place. For any two positive sequences $\{a_t\}_{t \ge 1}$ and $\{b_t\}_{t \ge 1}$, we write $a_t \preceq b_t$ if there exists some constant $C > 0$ such that $a_t \le C b_t$ for any t. The notation $a_t \preceq 1$ means $a_t = O(1)$.

Lemma 1 can thus be proven in a similar manner as Theorem 1 of Liu et al. (2018). Lemma 2 can be similarly proven as Lemma 1 of **?**. Theorem 2 can be proven in a similar manner as Theorem 3. In the following, we focus on proving Theorems 1, 3 and Lemma 3.

D.1. Proof of Lemma 3. We break the proof into three steps. In the first step, we use Berbee's coupling lemma (see Lemma 4.1 in **?**) to approximate $\sup_{f \in \mathcal{F}} |\sum_{t=0}^{T-1} f(Z_t)|$ by sum of i.i.d. variables. In the second step, we apply the tail inequality in Lemma 1 of **?** to bound the derivation between the empirical process and its mean. Finally, we apply the maximal inequality in Corollary 5.1 of Chernozhukov et al. (2014) to bound the expectation of the empirical process.

Step 1. Following the discussion below Lemma 4.1 of **?**, we can construct a sequence of random variables $\{Z_t^0 : t \geq 0\}$ such that

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| = \sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right|, \tag{15}$$

with probability at least $1 - T\beta(q)/q$, and that the sequences $\{U_{2i}^0 : i \geq 0\}$ and $\{U_{2i+1}^0 : i \geq 0\}$ are i.i.d. where $U_i^0 = (Z_{iq}^0, Z_{iq+1}^0, \cdots, Z_{iq+q-1}^0)$.

Recall that $\mathcal{I}_r = \{q\lfloor T/q\rfloor, q\lfloor T/q\rfloor + 1, \cdots, T-1\}$, we have

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/q\rfloor} f(Z_{tq+j}^0)\right| + \sup_{f\in\mathcal{F}}\left|\sum_{t\in\mathcal{I}_r} f(Z_t^0)\right|.$$

Under the boundedness assumption on F, the second term on RHS is bounded from above by Mq. Without loss of generality, suppose $\lfloor T/q\rfloor$ is an even number. The first term on the RHS can be bounded from above by $\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)|$. To summarize, we have shown

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| + Mq.$$

This together with (15) yields that

$$\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \leq \Pr\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) + \frac{T\beta(q)}{q}, \tag{16}$$

for any $\tau > 0$. By Bonferroni's inequality, we obtain

$$\Pr\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) \leq \sum_{j=0}^{2q-1}\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > \tau\right),$$

for any $\tau > 0$. Since the process is stationary, we obtain

$$\Pr\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) \leq 2q\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| > \tau\right).$$

Combining this together with (16) yields

$$\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \leq 2q\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| > \tau\right) + \frac{T\beta(q)}{q}. \tag{17}$$

By construction, $\{Z_{2tq}^0 : t \geq 0\}$ are i.i.d. This completes the proof of the first step.

Step 2. In the second step, we focus on relating the empirical process $\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)|$ to its expectation. Without loss of generality, assume $T = kq$ for some integer $k > 0$. Set the constants η and δ in Lemma 1 of **?** to 1, we obtain

$$\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| > 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| + \tau\right)$$
$$\leq 4\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + \exp\left(-\frac{\tau}{CM}\right),$$

for some constant $C > 0$. Combining this together with (17), we obtain

$$\begin{aligned}\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 4q\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| + 2\tau q + Mq\right) \\ \leq 8q\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + 2q\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned} \tag{18}$$

for any $\tau > 0$. This completes the proof of the second step.

Step 3. It remains to bound $\mathbb{E}\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)|$. By Corollary 5.1 of Chernozhukov et al. (2014), we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| \preceq \sqrt{\frac{\nu\sigma^2 T}{q}\log\left(\frac{AM}{\sigma}\right)} + \nu M\log\left(\frac{AM}{\sigma}\right).$$

Combining this together with (18), we obtain

$$\begin{aligned}\Pr\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > c\sqrt{\nu q\sigma^2 T\log\left(\frac{AM}{\sigma}\right)} + c\nu M\log\left(\frac{AM}{\sigma}\right) + cq\tau + Mq\right) \\ \leq Cq\exp\left(-\frac{\tau^2 q}{CT\sigma^2}\right) + Cq\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned}$$

for some constants $c, C > 0$ and any $\tau > 0, 1 \leq q < T/2$. The proof is hence completed.

D.2. Proof of Theorem 1. We introduce some notations. Let $\bar{Q}_{i,t} = \bar{Q}_i(A_{i,t}, m_i^a(\boldsymbol{A}_t), S_{i,t}, m_i^s(\boldsymbol{S}_t))$, $\bar{Q}_{i,t}(\boldsymbol{\pi}) = \bar{Q}_i(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t})$ and $\omega_{i,t}^* = \omega_i^*(\widetilde{S}_{i,t})$.

To prove Theorem 1, we apply the central limit theorem for mixing triangle arrays developed in **?**. Define $\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi})$ as

$$\frac{1}{N}\sum_{i=1}^{N}\left[V_i^*(\boldsymbol{\pi}) + \omega_{i,t}^*\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{R_{i,t} + \bar{Q}_{i,t+1}(\boldsymbol{\pi}) - \bar{Q}_{i,t} - V_i^*(\boldsymbol{\pi})\}\right],$$

we have $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi}) = T^{-1}\sum_{t=0}^{T-1}\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi})$. Under the stationarity assumption, σ^2 is equal to the variance of $\sqrt{N}\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi})$.

Suppose we have shown each $\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi})$ is an unbiased estimator for $V(\boldsymbol{\pi})$. For $t \in \{0, 1, \cdots, T-1\}$, let $x_t = (NT)^{-1/2}\{\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi}) - V(\boldsymbol{\pi})\}$. It suffices to show the conditions in (1)-(5) of **?** hold for $\{x_t : 0 \leq t < T\}$. We next verify these conditions.

Condition (1). Note that $\{R_{i,t}, \bar{Q}_i, \omega_i^*, V_i(\boldsymbol{\pi}) : 1 \leq i \leq N, t \geq 0\}$ are uniformly bounded from infinity, the set of functions $\{b_i : 1 \leq i \leq N\}$ are uniformly bounded from zero. As such, $\{x_t : 0 \leq t < T\}$ are uniformly bounded. Condition (1) thus holds for any $\nu^* > 0$.

Condition (2). This condition is automatically implied by the assumption that $NT\mathrm{Var}\{\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})\} \to \sigma^2 > 0$.

Condition (3). This condition holds by setting $\kappa = 0$ and $T_n = 0$ for any n.

Condition (4). Note that the strong mixing coefficients are upper bounded by the β-mixing coefficients. Under Condition (A4), we can take the sequence $\alpha(h)$ in Condition (4) by $\kappa_0\rho^h$.

Condition (5). Since $\kappa_0\rho^h$ decays to zero at an exponential rate as h grows to infinity, Condition (5) is automatically satisfied.

It remains to show $\mathbb{E}\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi}) = V(\boldsymbol{\pi})$ for any t. Suppose (A3) holds. Under the given conditions, we have $V_i^*(\boldsymbol{\pi}) = V_i(\boldsymbol{\pi})$. By Lemma 2, we have

$$\mathbb{E}\{R_{i,t} + \bar{Q}_{i,t+1}(\boldsymbol{\pi}) - \bar{Q}_{i,t} - V_i^*(\boldsymbol{\pi})|\boldsymbol{A}_t, \boldsymbol{S}_t\} = 0,$$

and hence,

$$\mathbb{E}\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{R_{i,t}+\bar{Q}_{i,t+1}(\boldsymbol{\pi})-\bar{Q}_{i,t}-V_i^*(\boldsymbol{\pi})\}=0.$$

Consequently, $\mathbb{E}\widehat{V}_t^{\mathrm{DR}*}(\boldsymbol{\pi})=N^{-1}\sum_{i=1}^{N}V_i(\boldsymbol{\pi})=V(\boldsymbol{\pi})$.

Suppose (A2) holds. Then we have $\omega_{i,t}^{*}=\omega_{i,t}$ for any i,t where $\omega_{i,t}$ is a shorthand for $\omega_i(\widetilde{S}_{i,t})$. As a result, for any i,t, the expectation of the density ratio $\omega_{i,t}^{*}\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))/b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})$ equals one. As such, we have

$$\begin{aligned}&\mathbb{E}\left\{V_i^*(\boldsymbol{\pi})-\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}V_i^*(\boldsymbol{\pi})\right\}\\&=V_i^*(\boldsymbol{\pi})\mathbb{E}\left\{1-\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\right\}=0.\end{aligned}\tag{19}$$

In addition, using similar arguments in (2), we have by (A2) that

$$\mathbb{E}\left\{\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}R_{i,t}\right\}=V_i(\boldsymbol{\pi}).\tag{20}$$

Moreover, by some calculations, we have

$$\begin{aligned}&\mathbb{E}\left\{\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\bar{Q}_{i,t}\right\}\\&=\mathbb{E}\left\{\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\bar{Q}_{i,t+1}(\boldsymbol{\pi})\right\}=\int_{\tilde{s}_i}Q_i^*(\pi_i(\boldsymbol{S}_{t+1}),\tilde{s}_i)p(\boldsymbol{\pi},\tilde{s}_i)d\tilde{s}_i.\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{\bar{Q}_{i,t+1}(\boldsymbol{\pi})-\bar{Q}_{i,t}\}\right]=0.$$

This together with (19) and (20) yields

$$\mathbb{E}\left[V_i^*(\boldsymbol{\pi})+\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{R_{i,t}+\bar{Q}_{i,t+1}(\boldsymbol{\pi})-\bar{Q}_{i,t}-V_i^*(\boldsymbol{\pi})\}\right]=V_i(\boldsymbol{\pi}).$$

It follows that $\mathbb{E}\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})=V(\boldsymbol{\pi})$.

Thus, $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$ is unbiased when either (A2) or (A3) holds. The proof is hence completed.

D.3. Proof of Theorem 3. Similarly, we define $\widehat{Q}_{i,t}$, $\widehat{Q}_{i,t}(\boldsymbol{\pi})$ and $\widehat{\omega}_{i,t}$ to be versions of $\bar{Q}_{i,t}$, $\bar{Q}_{i,t}(\boldsymbol{\pi})$ and $\omega_{i,t}^{*}$ with $\bar{Q}_i$ and ω_i^{*} replaced with $\widehat{Q}_i$ and $\widehat{\omega}_i$, respectively.

By Theorem 1, it suffices to show $\sqrt{NT}\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$ is asymptotically equivalent to $\sqrt{NT}\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$. Note that $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})-\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$ can be decomposed by $\eta_1+\eta_2+\eta_3+\eta_4+\eta_5$ where

$$\eta_1=\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left\{\omega_{i,t}^{*}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t),m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}\{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\},$$

$$\eta_2 = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\omega_{i,t}^*\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi})-\widehat{Q}_{i,t}-\bar{Q}_{i,t+1}(\boldsymbol{\pi})+\bar{Q}_{i,t}\},$$

$$\eta_3 = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t}^*)\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{R_{i,t}+\bar{Q}_{i,t+1}(\boldsymbol{\pi})-\bar{Q}_{i,t}-V_i^*(\boldsymbol{\pi})\},$$

$$\eta_4 = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t}^*)\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi})-\widehat{Q}_{i,t}-\bar{Q}_{i,t+1}(\boldsymbol{\pi})+\bar{Q}_{i,t}\},$$

$$\eta_5 = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t}^*)\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\}.$$

In the following, we show $|\eta_j| = o_p((NT)^{-1/2})$, for $j = 1, 2, \cdots, 5$.

Upper bounds on $|\eta_1|$: Note that $\eta_1 = N^{-1}\sum_{i=1}^N \eta_{1,i}$ where

$$\eta_{1,i} = \{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\omega_{i,t}^*\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}\right].$$

When (A2) holds, we have $\omega_{i,t}^* = \omega_{i,t}$ for any i, t. The expectation of the density ratio equals one. As a result, we have

$$\mathbb{E}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}=0,$$

for any i, t. In the following, we apply the Bernstein's inequality for exponential β-mixing processes (**?**) to bound $|\eta_1|$.

Under Condition (A4), the β-mixing coefficients of the sequence

$$\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1 : t\geq 0\right\}, \tag{21}$$

decays to zero at an exponential rate. In addition, all the terms in (21) are uniformly bounded by some constant $c > 0$. As a result,

$$\max_{t_1,t_2}\mathbb{E}\left|\omega_{i,t_1}\frac{\mathbb{I}(A_{i,t_1}=\pi_i(\boldsymbol{S}_t), \widetilde{A}_{i,t_1}=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t_1})}-1\right|\left|\omega_{i,t_2}\frac{\mathbb{I}(A_{i,t_2}=\pi_i(\boldsymbol{S}_t), \widetilde{A}_{i,t_2}=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t_2})}-1\right| = O(1).$$

It thus follows from Theorem 4.2 of **?** that there exists some constant $C > 0$ such that there exists some constant $C > 0$ such that for any $\tau \geq 0$ and integer $1 < q < T$,

$$\begin{aligned}&\max_i \Pr\left(\left|\sum_{t=0}^{T-1}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}\right|\geq 6\tau\right)\leq\frac{T}{q}\beta(q)\\&+\max_i\Pr\left(\left|\sum_{t\in\mathcal{I}_r}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}\right|\geq\tau\right)+4\exp\left\{-\frac{\tau^2}{Cq(T+\tau)}\right\},\end{aligned} \tag{22}$$

where $\mathcal{I}_r = \{q\lfloor T/q\rfloor, q\lfloor T/q\rfloor+1, \cdots, T-1\}$. Suppose $\tau \geq qc$. Notice that $|\mathcal{I}_r| \leq q$. It follows that

$$\max_i\Pr\left(\left|\sum_{t\in\mathcal{I}_r}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t)=m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}-1\right\}\right|\geq\tau\right)=0. \tag{23}$$

Under (A4), $\beta(q) = O(\rho^q)$. Set $q = -3\log(NT)/\log\rho$, we obtain $T\beta(q)/q = O(N^{-3}T^{-2})$. Set $\tau = \max\{2\sqrt{CqT\log(NT)}, 4Cq\log(NT)\}$, we obtain as $T \to \infty$ that

$$\frac{\tau^2}{2} \geq 2CqT\log(NT) \text{ and } \frac{\tau^2}{2} \geq 2Cq\tau\log(nT) \text{ and } \tau \geq qc.$$

Since $\sqrt{CqT\log(NT)} \gg 2Cq\log(NT)$, it follows from (22) and (23) that

$$\max_i \Pr\left(\left|\sum_{t=0}^{T-1}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})} - 1\right\}\right| \geq 12\sqrt{CqT\log(NT)}\right) \preceq N^{-2}T^{-2}.$$

By Bonferroni's inequality, we obtain the following event occurs with probability at least $1 - O(N^{-1}T^{-2})$,

$$\max_i \left|\sum_{t=0}^{T-1}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})} - 1\right\}\right| \leq 12\sqrt{CqT\log(NT)}.$$

It follows that

$$|\eta_1| \leq \frac{1}{N}\sum_{i=1}^N |\eta_{1,i}| \preceq \frac{\log(NT)}{\sqrt{T}}\left(\frac{1}{N}\sum_{i=1}^N |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})|\right), \tag{24}$$

with probability approaching 1. Under (A6) and the condition that $T \gg N\log^4(NT)$, we obtain $\eta_1 = o_p((NT)^{-1/2})$.

Upper bounds on $|\eta_2|$: When (A2) holds, we have $\omega_{i,t}^* = \omega_{i,t}$ for any i and t. As discussed in the proof of Theorem 1, we have $\mathbb{E}\eta_{2,i} = 0$ for any i where

$$\eta_{2,i} = \frac{1}{T}\sum_{t=0}^{T-1}\omega_{i,t}^*\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t} - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}\}.$$

In addition, notice that $\eta_{2,i}$ can be written as

$$\eta_{2,i} = \frac{1}{T}\sum_{t=0}^{T-1}\omega_{i,t}^*\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})\}.$$

We apply Lemma 3 to bound $\max_i |\eta_{2,i}|$. Define the class of functions $\mathcal{Q}_{i,\varepsilon}$ by

$$\left\{f \in \mathcal{Q} : \max_{i,a}\int_{\tilde{s}_i} |f(a,\tilde{s}_i) - Q_i^*(a,\tilde{s}_i)|^2 p_{i,b}(\tilde{s}_i)d\tilde{s}_i \leq \varepsilon\right\},$$

where $\varepsilon = \epsilon N^{-1/2}T^{-1/2}$ for some sufficiently small $\epsilon > 0$. It then follows from (A5)(ii) and (iii) that $\widehat{Q}_{i,\boldsymbol{\pi}} \in \mathcal{Q}_\varepsilon$ for any i with probability tending to 1. As such, we have

$$\eta_{2,i} \leq T^{-1}\sup_{Q_i \in \mathcal{Q}_{i,\varepsilon}}\left|\sum_{t=0}^{T-1}\omega_{i,t}^*\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\{f(\pi_i(\boldsymbol{S}_{t+1}), \widetilde{S}_{i,t+1}) - f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})\}\right|.$$

Consider the process $\{(\widetilde{S}_{i,t}, A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t+1}) : t \geq 0\}$. Under (A4), such a process has β-mixing coefficients $\{\beta^*(q) : q \geq 0\}$ that satisfies $\beta^*(q) = O(\rho^q)$ as well. For any f, define the function $g = g(f)$ such that

$$g(\widetilde{S}_{i,t}, A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t+1}, \pi_i(\boldsymbol{S}_t), \pi_i(\boldsymbol{S}_{t+1})) = \omega_{i,t}^*\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})} \times\{f(\pi_i(\boldsymbol{S}_{t+1}), \widetilde{S}_{i,t+1}) - f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})\},$$

almost surely. Consider the class of functions $\mathcal{G}_{i,\varepsilon} = \{g(f) : f \in \mathcal{Q}_{i,\varepsilon}\}$. Since $\mathcal{Q}_{i,\varepsilon}$ belongs to the class of VC-type class, so does $\mathcal{G}_{i,\varepsilon}$. Moreover, the VC-index of $\mathcal{G}_{i,\varepsilon}$ is the same as $\mathcal{Q}_{i,\varepsilon}$. Under the boundedness assumption in Theorem 2, we have

$$\mathbb{E}g^2(\widetilde{S}_{i,t}, A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t+1}, \pi_i(\boldsymbol{S}_t), \pi_i(\boldsymbol{S}_{t+1})) \le O(1)\varepsilon,$$

for some constant $O(1)$. In addition, the envelope function of $\mathcal{G}_{i,\varepsilon}$ is uniformly bounded.

Let $Z_{i,t} = (\widetilde{S}_{i,t}, A_{i,t}, m_i^a(\boldsymbol{A}_t), \widetilde{S}_{i,t+1}, \pi_i(\boldsymbol{S}_t), \pi_i(\boldsymbol{S}_{t+1}))$. Applying Lemma 3, we obtain

$$\begin{aligned}\max_i \Pr\left(\sup_{g\in\mathcal{G}_{i,\varepsilon}}\left|\sum_{t=0}^{T-1} g(Z_{i,t})\right| > c\sqrt{\nu q\varepsilon T\log\left(\frac{1}{\varepsilon}\right)} + c\nu\log\left(\frac{1}{\varepsilon}\right) + cq\tau + cq\right)\\ \le cq\exp\left(-\frac{\tau^2 q}{cT\varepsilon}\right) + cq\exp\left(-\frac{\tau}{c}\right) + \frac{T\beta(q)}{q},\end{aligned}$$

for some constant $c > 0$. Set $q = -2\log(NT)/\log\rho$, we have $T\beta(q)/q = O(N^{-2}T^{-1})$. Set $\tau = \max(2c\log(NT), \sqrt{2c\varepsilon T\log(NT)/q})$, the RHS is bounded by $O(N^{-2}T^{-1}\log(NT))$. By Bonferroni's inequality, we obtain with probability tending to 1 that

$$T|\eta_{2,i}| \le c\sqrt{\nu q\varepsilon T\log\left(\frac{1}{\varepsilon}\right)} + c\nu\log\left(\frac{1}{\varepsilon}\right) + cq\tau + cq, \quad \forall i \in \{1,\cdots,N\},$$

or equivalently,

$$\max_i |\eta_{2,i}| \preceq \sqrt{\frac{\epsilon}{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

under the condition that $T \gg N\nu^2\log^4(NT)$. Since ε can be chosen arbitrarily small, we obtain $\max_i |\eta_{2,i}| = o_p((NT)^{-1/2})$. This in turn implies $\eta_2 = o_p((NT)^{-1/2})$.

Upper bounds on $|\eta_3|$: Using similar arguments in proving $\eta_2 = o_p((NT)^{-1/2})$, we can show $\eta_3 = o_p((NT)^{-1/2})$. We omit the technical details to save space.

Upper bounds on $|\eta_4|$ and $|\eta_5|$: We show $\eta_4 = o_p((NT)^{-1/2})$ only. Using similar arguments, one can show $\eta_5 = o_p((NT)^{-1/2})$.

Note that

$$\begin{aligned}\eta_4 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t} - \omega_{i,t}^*)\frac{\mathbb{I}(A_{i,t} = \pi_i(\boldsymbol{S}_t), m_i^a(\boldsymbol{A}_t) = m_i^a(\boldsymbol{\pi}(\boldsymbol{S}_t)))}{b_i(\boldsymbol{\pi}|\widetilde{S}_{i,t})}\\ &\qquad\times\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})\}\\ &\le O(1)\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}|\widehat{\omega}_{i,t} - \omega_{i,t}^*||\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})|\\ &\le O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}[(\widehat{\omega}_{i,t} - \omega_{i,t}^*)^2 + \{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t+1}(\boldsymbol{\pi}) + \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2]\right\}\\ &\le O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t} - \omega_{i,t}^*)^2\right\} + O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\{\widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2\right\},\end{aligned}$$

where $O(1)$ denotes some universal constant, and the last two inequalities are due to Cauchy-Schwarz inequality.

To prove $\eta_4 = o_p((NT)^{-1/2})$, it suffices to show

$$\max_i \left[\frac{1}{T} \sum_{t=0}^{T-1} \{\widehat{Q}_{i,t}(\boldsymbol{\pi}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 \right] = o_p((NT)^{-1/2}), \tag{25}$$

and

$$\max_i \left\{ \frac{1}{T} \sum_{t=0}^{T-1} (\widehat{\omega}_{i,t} - \omega_{i,t}^*)^2 \right\} = o_p((NT)^{-1/2}). \tag{26}$$

The left-hand-side (LHS) of (25) can be upper bounded by

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \{f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 \right],$$

with probability tending to 1. Using similar arguments in proving $\eta_2 = o_p((NT)^{-1/2})$, we can show

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left| \frac{1}{T} \sum_{t=0}^{T-1} \{f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 - \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 \right| \preceq \frac{\epsilon}{\sqrt{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

with probability tending to 1. Under (A6), we have

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left| \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 \right| \preceq \frac{\epsilon}{\sqrt{NT}}.$$

It follows that

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left[\frac{1}{T} \sum_{t=0}^{T-1} \{f(\pi_i(\boldsymbol{S}_t), \widetilde{S}_{i,t}) - \bar{Q}_{i,t}(\boldsymbol{\pi})\}^2 \right] \preceq \frac{\epsilon}{\sqrt{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

with probability tending to 1. Let $\epsilon \to 0$, we obtain (25). Similarly, we can show (26) holds. The proof is hence completed.

APPENDIX E: ADDITIONAL SIMULATION RESULTS

We conduct paired two-sample t-test to test whether the MSE of DR is strictly smaller than that of QV in our simulation studies. The p-values of these t-tests are reported in Tables 1 and 2. It can be seen that most p-values are significant under the 0.05 significance level. Figure 10 plots the mean squared errors (MSEs) of the DR and QV estimators with a different y-axis scale. It can be clearly seen that our estimator achieves a strictly smaller MSE when $K = 6, 7$ or 9.

	$\sigma_R = 0$	$\sigma_R = 5$	$\sigma_R = 10$	$\sigma_R = 15$	$\sigma_R = 20$	$\sigma_R = 25$	$\sigma_R = 30$
$K = 9$	1.1×10^{-116}	1.3×10^{-102}	5.3×10^{-73}	8.4×10^{-40}	1.2×10^{-22}	2.8×10^{-14}	2.3×10^{-8}
$K = 8$	0.50	0.71	0.97	0.98	0.99	1.00	1.00
$K = 7$	6.6×10^{-87}	1.9×10^{-75}	3.4×10^{-41}	1.2×10^{-27}	4.9×10^{-19}	4.1×10^{-15}	6.4×10^{-7}
$K = 6$	8.8×10^{-43}	1.5×10^{-39}	1.2×10^{-19}	3.6×10^{-13}	8.8×10^{-10}	9.8×10^{-4}	5.8×10^{-3}

TABLE 1

P-values of paired two-sample t-test in the simulation study with different combinations of σ_R and K. T is fixed to 336.

	$\sigma_R = 0$	$\sigma_R = 5$	$\sigma_R = 10$	$\sigma_R = 15$	$\sigma_R = 20$	$\sigma_R = 25$	$\sigma_R = 30$
$K = 9$	1.1×10^{-116}	1.3×10^{-102}	5.3×10^{-73}	8.4×10^{-40}	1.2×10^{-22}	2.8×10^{-14}	2.3×10^{-8}
$K = 8$	0.50	0.71	0.97	0.98	0.99	1.00	1.00
$K = 7$	6.6×10^{-87}	1.9×10^{-75}	3.4×10^{-41}	1.2×10^{-27}	4.9×10^{-19}	4.1×10^{-15}	6.4×10^{-7}
$K = 6$	8.8×10^{-43}	1.5×10^{-39}	1.2×10^{-19}	3.6×10^{-13}	8.8×10^{-10}	9.8×10^{-4}	5.8×10^{-3}

TABLE 2

P-values of paired two-sample t-test in the simulation study with different combinations of T and K. σ_R is fixed to 15.

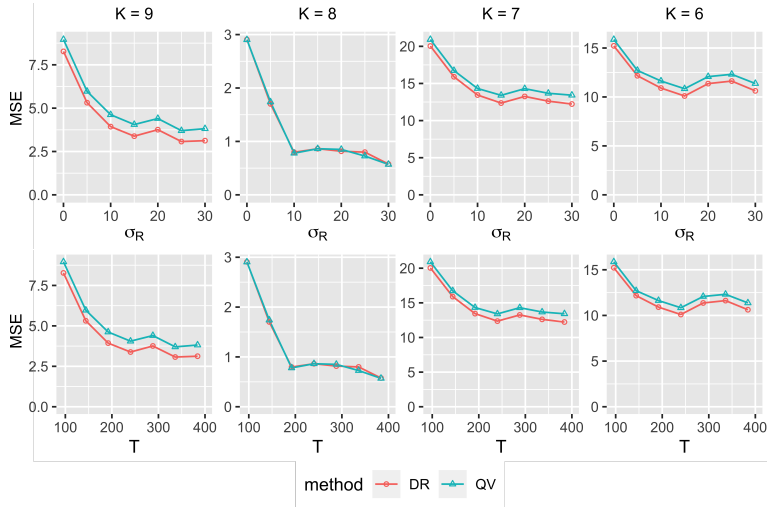

Fig 10: Mean squared errors of the DR and QV estimators, aggregated over 100 simulations. T is set to 336 (the experiment lasts for two weeks and each hour is treated as one time unit) in the top plots and σ_R is set to 15 in the bottom plots.