Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

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

Off-Policy Evaluation in Doubly Inhomogeneous Environments

Zeyu Bian, Chengchun Shi, Zhengling Qi & Lan Wang

To cite this article: Zeyu Bian, Chengchun Shi, Zhengling Qi & Lan Wang (11 Oct 2024): Off-Policy Evaluation in Doubly Inhomogeneous Environments, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2395593

To link to this article: https://doi.org/10.1080/01621459.2024.2395593

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2024.2395593

Off-Policy Evaluation in Doubly Inhomogeneous Environments

Zeyu Bian[a], Chengchun Shi[b], Zhengling Qi[c], and Lan Wang[a]

[a]Department of Management Science, University of Miami, Coral Gables, FL; [b]Department of Statistics, London School of Economics and Political Science, London, UK; [c]Department of Decision Sciences, The George Washington University, Washington, DC

ABSTRACT
This work aims to study off-policy evaluation (OPE) under scenarios where two key reinforcement learning (RL) assumptions—temporal stationarity and individual homogeneity are both violated. To handle the "double inhomogeneities", we propose a class of latent factor models for the reward and transition functions, under which we develop a general OPE framework that consists of both model-based and model-free approaches. To our knowledge, this is the first article that develops statistically sound OPE methods in offline RL with double inhomogeneities. It contributes to a deeper understanding of OPE in environments, where standard RL assumptions are not met, and provides several practical approaches in these settings. We establish the theoretical properties of the proposed value estimators and empirically show that our approach outperforms state-of-the-art methods. Finally, we illustrate our method on a dataset from the Medical Information Mart for Intensive Care. An R implementation of the proposed procedure is available at *https://github.com/ZeyuBian/2FEOPE*. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received June 2023
Accepted August 2024

KEYWORDS
Double inhomogeneities; Off-policy evaluation; Reinforcement learning; Two-way fixed effects model

1. Introduction

Reinforcement learning (RL, Sutton and Barto 2018) aims to optimize an agent's long-term reward by learning an optimal policy that determines the best action to take under every circumstance. RL is closely related to the dynamic treatment regimens (DTR) or adaptive treatment strategies in statistical research for precision medicine (Murphy 2003; Robins 2004; Qian and Murphy 2011; Kosorok and Moodie 2015; Shi et al. 2018; Tsiatis et al. 2019; Qi et al. 2020; Zhou, Zhu, and Qu 2024), which seeks to obtain the optimal treatment policy in finite horizon settings with a few treatment stages that maximizes patients' expected outcome. Nevertheless, statistical methods for DTR mentioned above normally cannot handle large or infinite horizon settings. They require the number of trajectories to tend to infinity to achieve estimation consistency, unlike RL, which works even with finite number of trajectories under certain conditions. In addition to precision medicine, RL has been applied to various fields, such as games (Silver et al. 2016), ridesharing (Xu et al. 2018), mobile health (Liao, Klasnja, and Murphy 2021) and robotics (Levine et al. 2020).

In this article, we focus on off-policy evaluation (OPE), whose objective is to evaluate the value function of a given target policy using data collected from a potentially different policy, known as the behavior policy. OPE is important in applications in which directly implementing a policy involves potential risks and high costs. For instance, in healthcare, it would be expensive to conduct a randomized experiment to recruit many individuals and follow them up for the duration of the entire experiment. Meanwhile, it might be unethical to directly apply a new treatment policy to some individuals without offline validation. It is therefore important to develop RL methods only using historical data, and OPE is particularly vital in offline RL. Generally speaking, existing OPE methods can be divided into four categories: model-based methods (Gottesman et al. 2019; Zhang et al. 2020), importance sampling methods (Precup 2000; Liu et al. 2018; Wang, Qi, and Wong 2021), direct methods (Luckett et al. 2020; Liao, Klasnja, and Murphy 2021; Shi et al. 2022), and doubly robust methods (Jiang and Li 2016; Uehara, Huang, and Jiang 2020; Kallus and Uehara 2022; Liao et al. 2022). See Uehara, Shi, and Kallus (2022) and the references therein for an overview.

Motivation. Most methods in the RL literature rely on the following two critical assumptions: temporal stationarity and individual homogeneity. The temporal stationarity assumption requires that the system dynamics for each subject do not depend on the time whereas individual homogeneity requires the system dynamics at each time to be identical across all subjects. Nonetheless, both conditions are likely to be violated in many RL applications, for example, mobile health and infectious disease control (Hu et al. 2022). This work draws partial motivation from the longitudinal data of septic patients obtained from the Medical Information Mart for Intensive Care (MIMIC-III, Johnson et al. 2016), a database containing information on critical care patients. Sepsis is a severe and potentially fatal condition that occurs when the human body's response to an infection injures its own tissues and organs (Singer et al. 2016). It can progress rapidly and cause multiple organ failures, resulting in an increased risk of death. Prompt treatment of sepsis is

CONTACT Chengchun Shi *C.Shi7@lse.ac.uk* Department of Statistics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

thus essential for improving patient outcomes and reducing mortality rates. However, the heterogeneity in patients' response to sepsis treatments (Evans et al. 2021), as well as a potentially nonstationary environment (the data includes patients' medical information over 10 years) make it challenging to effectively manage the illness using existing RL methods. Our analysis provides insights into evaluating the impact of different treatment strategies, facilitating the development of effective and personalized approaches to sepsis care.

In the statistical literature, Li et al. (2022) and Wang, Shi, and Wu (2023) developed hypothesis testing procedures to assess the stationarity assumption in RL, based on which policy learning procedures were proposed to handle possibly nonstationary environments. Chen, Song, and Jordan (2022) developed a transferred Q-learning algorithm and an auto-clustered policy iteration algorithm to handle heterogeneous data. However, these methods require either temporal stationarity or individual homogeneity, and would fail in doubly inhomogeneous environments when both assumptions are violated. Hu et al. (2022) proposed an algorithm to adaptively split the data into rectangles in which the system dynamics are identical over time and across individuals. They studied policy learning instead of OPE. In addition, they imposed a latent group structure over time and population. This structural assumption can be violated when the dynamics vary smoothly over both population and time.

Challenges. OPE is substantially more challenging under the doubly inhomogeneous environments. First, the evaluation target is different. In particular, most existing solutions developed in doubly homogeneous environments have predominantly focused on evaluating the expected long-term reward following the target policy aggregated over time and population. In contrast, the following four time- and/or individual-specific values are of particular interest in the presence of double inhomogeneities:

1. The expected long-term reward aggregated over both time and population;
2. The expected long-term reward aggregated over time for a given subject;
3. The expected reward at a given time aggregated over population;
4. The expected reward at a given time for a given subject.

Second, an unresolved challenge is how to efficiently borrow information over time and population for OPE. On one hand, to account for the subject heterogeneity or temporal nonstationarity, one could conduct OPE based on the data within each individual trajectory or at a given time. However, this approach may result in an estimator with a high variance. On the other hand, naively pulling data over population and time without careful considerations would lead to biased estimators.

Contributions. This work makes the following contributions. First, to our knowledge, it is the first study to investigate OPE in doubly inhomogeneous RL domains. Unlike prior works that primarily focused on evaluating the average effect over time and population, we provide a systematic approach for examining values that are specific to time and/or individuals. These values hold particular importance in the context of double inhomogeneities.

Second, we present a comprehensive framework for doubly inhomogeneous OPE which comprises both model-free and model-based methods. To effectively use information in the presence of temporal nonstationarity and individual heterogeneity, we introduce a class of two-way doubly inhomogeneous decision process (TWDIDP) models and develop corresponding policy value estimators. Our proposal shares similar spirits with the two-way fixed effects model that is widely studied in economics and social science (Angrist and Pischke 2009; Imai and Kim 2021). Nonetheless, our model is substantially more complicated due to the incorporation of carryover effects: in our model, the current treatment not only affects its immediate outcome, but also impacts the future outcomes through its effect on the future observation via the transition function. In contrast, the fixed effects models commonly employed in the panel data literature tend to exclude carryover effects (Imai and Kim 2019; Arkhangelsky et al. 2021).

Finally, we systematically investigate the theoretical properties of the proposed model-free method. In particular, we derive the convergence rates of various proposed value estimators, showing that the estimated average effect, individual-specific effect, time-specific effect and individual- and time-specific effect converge at a rate of $(NT)^{-1/2}$, $T^{-1/2}$, $N^{-1/2}$ and $\min^{-1/2}(N, T)$, respectively, up to some logarithmic factors, where N is the number of trajectories and T is the number of time points. We further establish the limiting distributions of these estimators.

Organization. The rest of this article is organized as follows. In Section 2, we introduce the proposed doubly inhomogeneous decision process model to incorporate temporal nonstationarity and individual heterogeneity. In Sections 3 and 4, we present our proposed model-free and model-based methods. We analyze their statistical properties in Section 5. A series of comprehensive simulation studies are conducted in Section 6. Finally, in Section 7, we illustrate the proposed approach using the MIMIC-III dataset.

2. Two-way Doubly Inhomogeneous Decision Processes

Data. We first describe the dataset. We assume the offline data consists of N independent trajectories, each with T many time points, and can be summarized as the following observation-action-reward triplets $\{(O_{i,t}, A_{i,t}, R_{i,t}) : 1 \leq i \leq N, 1 \leq t \leq T\}$ where i indexes the ith individual and t indexes the tth time point. For example, in mobile health applications, $O_{i,t} \in \mathbb{R}^d$ denotes the vector of covariates measured from the ith individual at time t where d is the dimension of the observation, $A_{i,t}$ denotes the treatment assigned to the ith individual at time t, and $R_{i,t} \in \mathbb{R}$ denotes the ith individual's clinical outcome at time t. Let $\mathcal{O}$ and $\mathcal{A}$ denote the observation and action space, respectively. We assume $\mathcal{A}$ is a discrete space whereas, $\mathcal{O}$ is a compact subspace of $\mathbb{R}^d$, and the reward is uniformly bounded. The bounded rewards assumption is commonly imposed in the RL literature (see e.g., Fan et al. 2020; Li et al. 2023).

Model. We next present the proposed two-way doubly inhomogeneous decision process model. In the RL literature, a common practice is to employ the Markov decision process (MDP, Puterman 2014) to model the data generating process. For simplicity, for now we assume both the observation and reward spaces are discrete, the MDP model essentially requires the reward and future observation to be conditionally independent of the past data history given the current observation-action pair so that the system dynamics are uniquely determined by the following Markov transition function p,

$$\mathbb{P}(O_{i,t+1} = o', R_{i,t} = r | A_{i,t} = a, O_{i,t} = o, \{O_{i,j}, A_{i,j}, R_{i,j}\}_{1 \le j < t}) = p(o', r | a, o), \tag{2.1}$$

which is assumed to be doubly homogeneous, that is, constant over time and population.

Instead of adopting the MDP model, we propose to use a more general model that relies on two key assumptions. First, we assume the existence of a set of individual- and time-specific latent factors $\{U_i\}_{i=1}^N$ and $\{V_t\}_{t=1}^T$ conditional on which the Markov assumption holds. More specifically, for any i and t, we assume

$$\begin{aligned}&\mathbb{P}(O_{i,t+1} = o', R_{i,t} = r | U_i = u_i, V_t = v_t, A_{i,t} = a, O_{i,t} = o, \\ &\{O_{i,j}, A_{i,j}, R_{i,j}, V_j\}_{1 \le j < t}) \\ &= p(o', r | u_i, v_t, a, o).\end{aligned} \tag{2.2}$$

Remark 1. Unlike (2.1), the transition function in (2.2) is both individual- and time-dependent due to the inclusion of U_i and V_t. The individual-specific factors can be viewed as certain individual baseline information (e.g., educational background) that does not vary over time whereas the time-specific factors correspond to certain external factors (e.g., holidays) that have common effects on all individuals.

Remark 2. Both $\{U_i\}_{i=1}^N$ and $\{V_t\}_{t=1}^T$ are unobserved in practice, leading to the violation of the Markov assumption. Indeed, the proposed data generating process can be viewed as a special class of partially observable MDPs (POMDPs, Sutton and Barto 2018) where the unobserved factors either do not evolve over time (e.g., $\{U_i\}_{i=1}^N$) or do not vary across individuals (e.g., $\{V_t\}_{t=1}^T$). More generally, one may allow the latent factors to evolve over both time and population. However, this makes the subsequent policy evaluation extremely challenging. In contrast, our proposal decomposes these factors into individual-only and time-only effects, which can be consistently estimated when both N and T diverge to infinity. Such latent factor models are widely used in finance (Ross 1976), economics (Bai and Ng 2002) and psychology (Bollen 2002).

Remark 3. The assumption of discrete rewards and observations is used merely to simplify the presentation. Our proposed methodology can be equally applied to scenarios with continuous observation/reward spaces as well. In these cases, we use $p(\bullet, \bullet | U_i, V_t, A_{i,t}, O_{i,t})$ to represent the conditional density function of $(O_{i,t+1}, R_{i,t})$ given U_i, V_t, $A_{i,t}$, and $O_{i,t}$.

Second, we impose an additivity assumption, which requires the transition function p to be additive in u, v and (a, o), i.e.,

$$\begin{aligned}p(o', r | u_i, v_t, a, o) = \pi_u p_{u_i}(o', r | u_i) \\ + \pi_v p_{v_t}(o', r | v_t) + \pi_0 p_0(o', r | a, o)\end{aligned} \tag{2.3}$$

for some nonnegative constants π_u, π_v, and π_0 that satisfy $\pi_u + \pi_v + \pi_0 = 1$ and some unknown conditional probability density (mass) functions p_u, p_v, and p_0.

The additivity assumption in (2.3) essentially assumes that the transition function corresponds to a mixture of p_u, p_v, and p_0, with the mixing weights given by π_u, π_v, and π_0, respectively. Under the additivity assumption, p_u and p_v correspond to the individual- and time-specific effects, respectively, and are independent of the current observation-action pair. The function p_0 corresponds to the main effect shared over time and subjects. Meanwhile, such an additivity assumption can be further relaxed; see Section 3.2 for details.

Multiplying both sides of (2.3) by r and integrating with respect to r and o', we obtain

$$R_{i,t} = \theta_i + \lambda_t + r_1(a, o) + \varepsilon_{i,t}, \tag{2.4}$$

where $\theta_i = \pi_u \int r p_{u_i}(o', r | u_i) dr do'$, $\lambda_t = \pi_v \int r p_{v_t}(o', r | v_t) dr do'$, $r_1(a, o) = \pi_0 \int r p_0(o', r | a, o) dr do'$, and $\varepsilon_{i,t} = R_{i,t} - \mathbb{E}(R_{i,t} | A_{i,t} = a, O_{i,t} = o)$ has conditional mean zero. Models of this type are referred to as the two-way fixed-effects (2FE) model in the panel data literature (see e.g., Imai and Kim 2021). Nonetheless, our model allows the current treatment to not only affect the immediate outcome, but also impact the future outcomes through its effect on the future observations via the transition function in (2.3).

Remark 4. Our additivity assumption (2.3) is motivated by the increased popularity of the fixed-effect models in the panel data literature, due to its interpretability and the ability to account for unobserved variables. As commented by Green, Kim, and Yoon (2001), fixed effects regression can scarcely be faulted for being the bearer of bad tidings. Such a model has emerged as a crucial tool assisting researchers in various fields such as medical and political science, facilitating the derivation of scientific conclusions (Hotz and Xiao 2011; Bachhuber et al. 2014; Dwivedi et al. 2022).

To summarize the data generating process, the latent factors $\{U_i\}_i$ and $\{V_t\}_t$ are sampled prior to all interactions with the environment. For a specific trajectory i, at each time point t, we observe $O_{i,t}$ according to the transition model (2.3). Next, the agent takes an action $A_{i,t}$ according to the observed data history and receives an immediate reward $R_{i,t}$ according to (2.3). Finally, the environment transits into the next state, yielding $O_{i,t+1}$. A causal diagram illustrating the data generating procedure can be found in Figure 1 of the supplementary article. In what follows, we assume the latent factors $\{U_i\}_i$ and $\{V_t\}_t$ are fixed and use $\{u_i\}_i$ and $\{v_t\}_t$ to denote their realizations. Other random variables in the environment will not alter their values. In the sequel, all the expectations mentioned are implicitly conditional on $\{U_i\}_i$ and $\{V_t\}_t$.

Estimands. Finally, we define our target estimand of interest. A policy prescribes how an agent acts and makes decisions. Mathematically, it maps the space of observed data history to a probability mass function on the action space, representing the probability that a given individual receives a given treatment at each time point. Throughout this article, we focus on evaluating *stationary* policies where the action selection probability depends on history only through the current observation and this dependence is stationary over time. More specifically, following a given stationary policy π, the ith individual will receive treatment a with probability $\pi(a|O_{i,t})$. Meanwhile, the proposed method can be extended to evaluate possibly history-dependent policies; see Section A.2 of the supplementary article.

For a target policy π, we define the following four estimands of interest: (i) the average effect $\eta^\pi := (NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T \mathbb{E}^\pi(R_{i,t})$; (ii) the individual-specific effect given the observed initial observation $\eta_i^\pi := T^{-1}\sum_{t=1}^T \mathbb{E}^\pi(R_{i,t}|O_{i,1})$; (iii) the time-specific effect $\eta_t^\pi := N^{-1}\sum_{i=1}^N \mathbb{E}^\pi(R_{i,t})$ and (iv) the individual- and time-specific effect $\eta_{i,t}^\pi := \mathbb{E}^\pi(R_{i,t}|O_{i,1})$. Here, $\mathbb{E}^\pi$ means that the expectation is taken by assuming the system dynamics follow the target policy π. In defining η_i^π and $\eta_{i,t}^\pi$, we include $O_{i,1}$ in the conditioning set to eliminate their variability resulting from marginalizing over the initial observation distribution. This is reasonable, as the initial observation distribution may no longer be identical across different subjects due to individual heterogeneity, making it impossible to infer consistently from the data. We focus on estimating (iv) $\eta_{i,t}^\pi$ in the next two sections, based on which estimators for (i)–(iii) can be easily derived by taking the average over time and/or population.

3. Model-free OPE

We now develop model-free methodologies to learn $\eta_{i,t}^\pi$: the ith subject's average reward at time t given $O_{i,1}$. Model-free methods construct the estimator without directly learning the transition function. Compared to model-based methods which directly learn the transition function to derive the estimator, they are preferred in settings with a large observation space, or where the transition function is highly complicated and can be misspecified. In RL, both model-free and model-based methods have their own unique strengths, and we discuss this point thoroughly in Section A.3 of the supplementary article.

Challenge. Before presenting our proposal, we outline the challenges in consistently estimating the policy value. First, existing model-free methods developed in the RL literature (see e.g., Luckett et al. 2020; Shi et al. 2022) focused on learning the long-term reward in a stationary environment. These methods are not applicable to learn the expected reward at a given time with nonstationary transition functions. Second, in the DTR literature, backward induction or dynamic programming is widely employed to evaluate the value function in the sparse reward setting where the reward is obtained at the last stage and all the immediate rewards equal zero. It is applicable to evaluate $\mathbb{E}^\pi(R_{i,t})$ in nonstationary environments. Nonetheless, it requires all individual trajectories to follow the same distribution and is thus inapplicable to our setting.

Q-function. Our proposal extends the backward induction to the doubly inhomogeneous environments. We first define the following individual- and time-specific Q-function

$$Q_{i,t_1,t_2}^\pi(o,a) = \mathbb{E}^\pi(R_{i,t_2}|A_{i,t_1}=a, O_{i,t_1}=o), \tag{3.1}$$

for any $1 \le i \le N$ and $1 \le t_1 \le t_2 \le T$. To elaborate on this definition, we consider two particular choices of t_1. First, when $t_1 = t_2$, (3.1) reduces to the conditional mean of R_{i,t_2} given (A_{i,t_2}, O_{i,t_2}), which equals $\theta_i + \lambda_t + r_1(A_{i,t}, O_{i,t})$ (see (2.4)) under additivity. Second, when $t_1 = 1$, it is immediate to see that

$$\eta_{i,t_2}^\pi = \sum_a Q_{i,1,t_2}^\pi(O_{i,1}, a)\pi(a|O_{i,1}). \tag{3.2}$$

As such, it suffices to learn $Q_{i,1,t}^\pi$ to construct estimators for $\eta_{i,t}^\pi$.

Remark 5. In the RL literature, the Q-function is typically defined as the cumulative reward starting from a given time t_1. Our Q-function in (3.1) differs in that: (i) it is individual-specific where the subscript i encodes its dependence upon the latent factor u_i; (ii) it is the conditional mean of the immediate reward at time t_2 only instead of the cumulative reward since our objective here lies in evaluating $\mathbb{E}^\pi(R_{i,t_2})$.

Backward induction. We propose to use backward induction to compute an estimated Q-function $\widehat{Q}_{i,1,t}^\pi$ for $Q_{i,1,t}^\pi$ and then plug this estimator into (3.2) to construct the policy value estimator. To begin with, consider the reward function $\{Q_{i,t,t}^\pi\}_{i,t}$. As shown in (2.4), under the two-way fixed-effect model, we have $Q_{i,t,t}^\pi(o,a) = r_1(o,a) + \theta_i + \lambda_t$ for any i and t. This motivates us to consider the following optimization problem:

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}}, \widehat{r}_1) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\lambda},r_1} \sum_{i,t}[R_{i,t} - \theta_i - \lambda_t - r_1(O_{i,t}, A_{i,t})]^2, \tag{3.3}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^\top \in \mathbb{R}^N$, and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_T)^\top \in \mathbb{R}^T$. To guarantee the uniqueness of the solution to (3.3), we impose the identifiability constraints $\sum_i \theta_i = \sum_t \lambda_t = 0$. There are other constraints one could consider, but they all lead to the same final estimators.

We next estimate $\{Q_{i,t-1,t}\}_{i,t}$. According to the Bellman equation, we obtain

$$\begin{aligned} &Q_{i,t-1,t}(O_{i,t-1}, A_{i,t-1}) \\ &= \mathbb{E}\Big[\sum_a \pi(a|O_{i,t})Q_{i,t,t}(O_{i,t}, a)\big|A_{i,t-1}, O_{i,t-1}\Big]. \end{aligned}$$

Under the additivity assumption, we can similarly obtain a two-way decomposition for $Q_{i,t-1,t}$; see Proposition 1 for a formal statement. This allows us to solve a constrained optimization problem similar to (3.3) to estimate $Q_{i,t-1,t}$. We next repeat this procedure to recursively estimate $\{Q_{i,t-2,t}\}_{i,t}$, $\{Q_{i,t-3,t}\}_{i,t}$, $\cdots$, $\{Q_{i,1,t}\}_{i,t}$ based on the Bellman equation and finally construct the policy value estimator via (3.2). We summarize our estimating procedure in Algorithm 1. The following proposition formally states the two-way structure of these Q-functions.

Proposition 1. For any integer $k < t$, the Q-function $Q_{i,t-k+1,t}^\pi(o,a)$ satisfies

$$Q_{i,t-k+1,t}^\pi(o,a) = r_k^\pi(o,a) + \theta_{k,i}^\pi + \lambda_{k,t}^\pi,$$

where $\theta_{k,i}^\pi$ and $\lambda_{k,t}^\pi$ depend only on i, k, π and t, k, π, respectively.

In what follows, we omit π in $r_k^\pi(o,a)$, $\theta_{k,i}^\pi$, and $\lambda_{k,t}^\pi$ when there is no confusion.

To conclude this section, we draw a comparison with the classical backward induction in the DTR literature (Murphy 2003; Robins 2004). First, the classical backward induction algorithm aims to learn the Q-function under an optimal policy and

Algorithm 1 Pseudocode for Estimating η^{π}_{i,t^*}.

function $(R_{i,t}, A_{i,t}, O_{i,t}$ for $1 \le i \le N, 1 \le t \le T)$
  Set iteration counter $k \leftarrow 1$.
  Solve $(\widehat{\theta}_i, \widehat{\lambda}_t, \widehat{r}_1) = \arg\min_{\theta_i,\lambda_t,r_1} \sum_{i,t}[R_{i,t} - \theta_i - \lambda_t - r_1(O_{i,t}, A_{i,t})]^2$.
  Compute $\widehat{Q}^{\pi}_{i,t,t}(O_{i,t}, a) = \widehat{r}_1(O_{i,t}, a) + \widehat{\theta}_i + \widehat{\lambda}_t$, for all i, t, and $a \in \mathcal{A}$.
  repeat
    $k \leftarrow k + 1$
    For all i and $t \ge k$, solve

$$\min_{\theta_{k,i},\lambda_{k,t},r_k} \sum_{i,t} \left[\sum_{a\in\mathcal{A}} \pi(a|O_{i,t-k+2}) \widehat{Q}^{\pi}_{i,t-k+2,t}(O_{i,t-k+2}, a) - \theta_{k,i} - \lambda_{k,t} - r_k(O_{i,t-k+1}, A_{i,t-k+1}) \right]^2 .$$

    Compute $\widehat{Q}^{\pi}_{i,t-k+1,t}(O_{i,t-k+1}, a) = \widehat{r}_k(O_{i,t-k+1}, a) + \widehat{\theta}_{k,i} + \widehat{\lambda}_{k,t}$, for $a \in \mathcal{A}$.
  until $k = t^*$
  $\widehat{\eta}^{\pi}_{i,t^*} = \sum_a \widehat{Q}^{\pi}_{i,1,t^*}(O_{i,1}, a)\pi(a|O_{i,1})$

derive the optimal policy as the greedy policy with respect to the estimated Q-function. To the contrary, the proposed algorithm learns the Q-function under a fixed target policy for the purpose of policy evaluation. Second, classical backward induction requires the computation of the Q-function recursively from time t till the beginning. However, it is worth mentioning that our estimated Q-function converges exponentially fast to a constant function with respect to the lag k (see Section B.2.2 of the supplementary material). We refer to this phenomenon as Q-function degeneracy. As such, early stopping can be potentially employed in Algorithm 1 to speed up the computation.

Finally, the proposed backward induction allows us to to efficiently borrow information under the additivity assumption. Specifically, during each iteration, we pull all the relevant data together to estimate the Q-function. This allows us to consistently estimate the main effect (shared by all observations) at a rate of $(NT)^{-\alpha}$ which depends on both N and T, and the exponent $0 < \alpha \le 1/2$ depends on the nonparametric methods being used to solve the optimization problem. Meanwhile, the two-way fixed effects θs and λs converge at $T^{-1/2}$ and $N^{-1/2}$, respectively, up to some logarithmic factors. To the contrary, the estimator obtained via the classical backward induction typically converges at a rate of $N^{-\alpha'}$ for some $0 < \alpha' \le 1/2$ in individual-homogeneous and history-dependent[1] environments.

3.1. A Linear Sieve Estimator for Two-Way Fixed Effects Model

Notation. Given arbitrary $\{x_{i,t}\}_{1\le i\le N,1\le t\le T}$, let $\boldsymbol{x} \in \mathbb{R}^{NT}$ denote the vector whose $((t-1)N + i)$th element equals $x_{i,t}$. That is, $\boldsymbol{x}$ is constructed by stacking the N elements at the first time point, followed by the N elements at the second time point, and continuing in this manner until the N elements from the final time point are included, that is,

$$\boldsymbol{x} = (x_{1,1}, x_{2,1}, \ldots, x_{N,1}, x_{1,2}, \ldots, x_{N-1,T}, x_{N,T})^\top.$$

Similarly, given a set of vectors $\{\boldsymbol{x}_{i,t}\}_{i,t}$, let $\boldsymbol{X}$ denote the matrix whose $((t-1)N+i)$th row equals $\boldsymbol{x}_{i,t}$. To implement Algorithm 1, we need to solve two-way fixed effects models repeatedly for value function estimation. To simplify the presentation, we focus on the estimation of $Q^{\pi}_{i,t,t}(O_{i,t}, a_{i,t})$ (see (3.3)). We propose to approximate the main effect function $r_1(o, a)$ using linear sieves (Huang 1998; Chen and Christensen 2015). Under mild conditions, there exists a set of vectors $\{\boldsymbol{\beta}^*_a\}$ such that the approximation error is negligible, that is, $\sup_{o,a} |r_1(o, a) - \boldsymbol{\Phi}_L(o)^\top \boldsymbol{\beta}^*_a| = O(L^{-p/d})$, where $\boldsymbol{\Phi}_L(o)$ is a vector consisting of L sieve basis functions, for example, splines or wavelet bases, and $p > 0$ measures the smoothness of the system dynamics; see Section B.2.1 for more details. For simplicity, we now focus on the binary action space setting, in which $\mathcal{A} = \{0, 1\}$.

The two-way fixed effects model in (2.3) can be represented in the following matrix form: $\boldsymbol{R} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{M} + \boldsymbol{\varepsilon}$, where $\boldsymbol{R} = (R_{1,1}, R_{2,1}, \ldots, R_{N,1}, R_{1,2}, \ldots, R_{N-1,T}, R_{N,T})^\top \in \mathbb{R}^{NT}$, $\boldsymbol{\alpha} = (\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top)^\top$, $\boldsymbol{B} = (\mathbf{1}_T \otimes \boldsymbol{I}_N, \boldsymbol{I}_T \otimes \mathbf{1}_N) \in \mathbb{R}^{NT\times(N+T)}$ is the design matrix, $\boldsymbol{I}_N$ is a $N \times N$ identity matrix, $\mathbf{1}_T$ is a vector of length T with all elements one, $\boldsymbol{M} = (r_1(O_{1,1}, A_{1,1}), r_1(O_{2,1}, A_{2,1}), \ldots, r_1(O_{N,1}, A_{N,1}), r_1(O_{1,2}, A_{1,2}), \cdots, r_1(O_{N,T}, A_{N,T}))^\top \in \mathbb{R}^{NT}$, and $\otimes$ is the Kronecker product. In what follows, we will omit the indices of these matrices and vectors when there is no confusion. Let $\boldsymbol{\Phi}_{i,t} = ((1 - A_{i,t})\boldsymbol{\Phi}_L^\top(O_{i,t}), A_{i,t}\boldsymbol{\Phi}_L^\top(O_{i,t}))^\top$, and let $\boldsymbol{\Phi}$ be the $\mathbb{R}^{NT\times 2L}$ matrix

$$(\boldsymbol{\Phi}_{1,1}^\top, \boldsymbol{\Phi}_{2,1}^\top, \ldots, \boldsymbol{\Phi}_{N,1}^\top, \boldsymbol{\Phi}_{1,2}^\top, \ldots, \boldsymbol{\Phi}_{N-1,T}^\top, \boldsymbol{\Phi}_{N,T}^\top)^\top.$$

By the Frisch–Waugh–Lovell theorem (Frisch and Waugh 1933; Lovell 1963), a closed-form estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top)^\top$ can be obtained accordingly:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Phi}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{R}, \tag{3.4}$$

where $\boldsymbol{P}$ is the projection matrix: $\boldsymbol{P} = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{+}\boldsymbol{B}^\top$, and $(\boldsymbol{B}^\top\boldsymbol{B})^{+}$ is the Moore–Penrose inverse of the matrix $\boldsymbol{B}^\top\boldsymbol{B}$. Given $\widehat{\boldsymbol{\beta}}$, the estimator for the main effect function $r_1(O_{i,t}, A_{i,t})$ (denoted by $\widehat{r}_1$) can then be obtained, based on which the fixed effects can be estimated. Specifically, under the constraints that $\sum_{i=1}^N \theta_i = \sum_{t=1}^T \lambda_t = 0$, we have

$$\widehat{\theta}_i = T^{-1}\sum_{t=1}^{T}(R_{i,t} - \widehat{r}_1(O_{i,t}, A_{i,t})), \text{ and}$$

$$\widehat{\lambda}_t = N^{-1}\sum_{i=1}^{N}(R_{i,t} - \widehat{r}_1(O_{i,t}, A_{i,t})).$$

The resulting estimated Q-function is given by $\widehat{Q}^{\pi}_{i,t,t}(o, a) = \widehat{\theta}_i + \widehat{\lambda}_t + \boldsymbol{\Phi}_L(o)^\top\widehat{\boldsymbol{\beta}}_a$.

[1] The transition function depends on the entire history instead of the current observation-action pair.

3.2. Beyond the Additivity Assumption

In this section, we discuss two extensions of the additivity assumption (2.3). This first extension allows the time- and individual-specific latent factors to additionally depend on the action, whereas the second extension permits the system dynamics to incorporate interactive effects between time and individual.

Action-dependent time- and individual-specific effects. We first consider the following relaxation of the additivity assumption:

$$p(o', r|u_i, v_t, a, o) = \pi_u p_{u_i,a}(o', r|u_i) + \pi_v p_{v_t,a}(o', r|v_t) \\ + \pi_0 p_0(o', r|a, o),$$

such that $\pi_u + \pi_v + \pi_0 = 1$. That is, we now allow both the $p_{u_i}(o', r|u_i)$ and $p_{v_t}(o', r|v_t)$ in (2.2) to depend on the action a, and now the Q-function $Q^{\pi}_{i,t-k+1,t}(o, a)$ satisfies

$$Q^{\pi}_{i,t-k+1,t}(o, a) = r^{\pi}_k(o, a) + \theta^{\pi}_{k,i}(a) + \lambda^{\pi}_{k,t}(a),$$

where $\theta^{\pi}_{k,i}(a)$ and $\lambda^{\pi}_{k,t}(a)$ are action-dependent. Since a is binary, each iteration requires to estimate $2N + 2T_k$ fixed effects. The proposed approach can be easily extended to solve this new problem without extra complications. We omit the details to save space.

Interactive time- and individual-specific effects. The second extension is motivated by the factor model, which is extensively used in the panel data literature to relax the additivity assumption (Bai and Ng 2002). In our setup, consider the reward regression model in (2.4). The factor model replaces the additive terms $\theta_i + \lambda_t$ in (2.4) with an interaction term $\boldsymbol{\gamma}_i^{\top}\boldsymbol{\alpha}_t$, resulting in

$$R_{i,t} = \boldsymbol{\gamma}_i^{\top}\boldsymbol{\alpha}_t + r_1(A_{i,t}, O_{i,t}) + \varepsilon_{i,t}, \tag{3.5}$$

where $\boldsymbol{\gamma}_i \in \mathbb{R}^h$ and $\boldsymbol{\alpha}_t \in \mathbb{R}^h$ denote the vectors of unobserved common time- and individual-specific factors, respectively. By definition, it covers the additive model as a special case by setting $h = 2$, $\boldsymbol{\gamma}_i = (1, \theta_i)^{\top}$ and $\boldsymbol{\alpha}_t = (\lambda_t, 1)^{\top}$.

Combining (3.5) together with a completeness assumption, which requires functions in the form of the right-hand-side of (3.5) to be closed under the Bellman operator (see Assumption 2 in Section 5), we can show that $Q_{i,t-k+1,t}$ maintains a factor structure for any $k < t$. Similar to Algorithm 1, backward induction remains applicable for estimating $\eta^{\pi}_{i,t}$.

Finally, we provide a model diagnostic procedure in Section A.1 in the supplementary article to assess the additivity assumption. Specifically, we tackle the model selection problem of determining whether the additive or interactive model better fits the data. Therein, we apply this procedure across various of synthetic environments to demonstrate its effectiveness.

4. Model-based OPE

In this section, we develop model-based methods that derive the off-policy value estimator by learning the system dynamics. Recall that under the additivity assumption,

$$R_{i,t} = \theta_i + \lambda_t + r_1(O_{i,t}, A_{i,t}) + \varepsilon_{i,t}.$$

As we discussed in Section 3, the main effect r_1, as well as the individual- and time-specific effects can be estimated by solving the following optimization problem,

$$\arg\min_{\{\theta_i\}_i, \{\lambda_t\}_t, r_1} \sum_{i,t} [R_{i,t} - \theta_i - \lambda_t - r_1(O_{i,t}, A_{i,t})]^2.$$

In addition, we need to estimate the mixing probabilities π_u, π_v, π_0 as well as the distribution functions $p_{u_i}(o'|u_i)$, $p_{v_t}(o'|v_t)$, $p_0(o'|a, o)$, obtained by marginalizing over $p_{u_i}(o', r|u_i)$, $p_{v_t}(o', r|v_t)$, $p_0(o', r|a, o)$ in (2.3).

Given these estimators, we employ a simulation-based method to construct the policy value. To be more specific, based on the estimated transition function, we simulate an observation $O^*_{i,2}$ based on the observed $O_{i,1}$ under the target policy π. We next sequentially simulate a sequence of observations $\{O^*_{i,t}\}_t$ under π and compute the estimated reward $\pi(a|O^*_{i,t})(\widehat{r_1}(O^*_{i,t}, a) + \widehat{\theta}_i + \widehat{\lambda}_t)$. Finally, we repeat this procedure sufficiently many times and average all the estimated rewards across different simulations.

Likelihood. It remains to estimate p_u, p_v, p_0 and π_u, π_v, π_0. Given the latent factors, the likelihood function is proportional to the following,

$$\begin{aligned} &\prod_{i=1}^{N}\prod_{t=2}^{T} p(O_{i,t}|u_i, v_{t-1}, A_{i,t-1}, O_{i,t-1}; \boldsymbol{\Theta}) \\ &= \prod_{i=1}^{N}\prod_{t=2}^{T} [\pi_u p_{u_i}(O_{i,t}|u_i; \boldsymbol{\Theta}_u) + \pi_v p_{v_t}(O_{i,t}|v_{t-1}, \boldsymbol{\Theta}_v) \\ &\quad + \pi_0 p_0(O_{i,t}|A_{i,t-1}, O_{i,t-1}; \boldsymbol{\Theta}_0)], \end{aligned} \tag{4.1}$$

where we parameterize the transition model by $\boldsymbol{\Theta} = \{\pi_0, \pi_u, \pi_v, \boldsymbol{\Theta}_0, \boldsymbol{\Theta}_v, \boldsymbol{\Theta}_u\}$, and $\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_v, \boldsymbol{\Theta}_u$ are the parameters corresponding to models p_0, p_v, and p_u, respectively.

We introduce a latent variable $Z_{i,t} \in \{0, 1, 2\}$ such that $Z_{i,t} = 0$, if $O_{i,t}$ is generated by p_0, $Z_{i,t} = 1$, if $O_{i,t}$ is generated by p_v, and $Z_{i,t} = 2$ otherwise. Directly maximizing the likelihood in (4.1) is challenging, since it requires to marginalize over $Z_{i,t}$. We propose to use the EM (Dempster, Laird, and Rubin 1977) algorithm for parameter estimation. The EM algorithm recursively alternates between an E-step for computing conditional expectation and an M-step for maximizing the likelihood. We detail the two steps below.

E-step. Similar to (4.1), the complete log-likelihood involving $\{O_{i,t}\}_{i,t}$ and $\{Z_{i,t}\}_{i,t}$ is

$$\begin{aligned} &l(\boldsymbol{O}, \boldsymbol{Z}|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}; \boldsymbol{\Theta}) \\ &\propto \sum_{i=1}^{N}\sum_{t=2}^{T} \log p(O_{i,t}|Z_{i,t}, u_i, v_{t-1}, A_{i,t-1}, O_{i,t-1}; \boldsymbol{\Theta}) p(Z_{i,t}; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{N}\sum_{t=2}^{T} [\mathbb{I}(Z_{i,t} = 0) \log(\pi_0 p_0(O_{i,t}|A_{i,t-1}, O_{i,t-1}; \boldsymbol{\Theta}_0)) \\ &\quad + \mathbb{I}(Z_{i,t} = 1) \log(\pi_v p_{v_t}(O_{i,t}|v_t; \boldsymbol{\Theta}_v)) \\ &\quad + \mathbb{I}(Z_{i,t} = 2) \log(\pi_u p_{u_i}(O_{i,t}|u_i; \boldsymbol{\Theta}_u))]. \end{aligned} \tag{4.2}$$

Given a current estimate of $\boldsymbol{\Theta}$, say $\widetilde{\boldsymbol{\Theta}}$, define $\Gamma(\boldsymbol{\Theta}|\widetilde{\boldsymbol{\Theta}})$ as the expected value of $l(\boldsymbol{O}, \boldsymbol{Z}|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}; \boldsymbol{\Theta})$ with respect to the currently estimated conditional distribution function $p(\boldsymbol{Z}|\boldsymbol{U},\boldsymbol{V},\boldsymbol{O},\boldsymbol{A};\widetilde{\boldsymbol{\Theta}})$. We aim to calculate Γ in this step. It follows from (4.2) that

$$\Gamma(\boldsymbol{\Theta}|\widetilde{\boldsymbol{\Theta}}) = \sum_{i=1}^{N}\sum_{t=2}^{T}[p(Z_{i,t}=0|O_{i,t},A_{i,t-1},O_{i,t-1};\widetilde{\boldsymbol{\Theta}}_0) \log(\pi_0 p_0(O_{i,t}|A_{i,t-1},O_{i,t-1};\boldsymbol{\Theta}_0)) + p(Z_{i,t}=1|O_{i,t},v_t;\widetilde{\boldsymbol{\Theta}}_v)\log(\pi_v p_{v_t}(O_{i,t}|v_t;\boldsymbol{\Theta}_v)) + p(Z_{i,t}=2|O_{i,t},u_i;\widetilde{\boldsymbol{\Theta}}_u)\log(\pi_u p_{u_i}(O_{i,t}|u_i;\boldsymbol{\Theta}_u))]. \tag{4.3}$$

M-step. We aim to update the model parameter $\boldsymbol{\Theta}_{new}$ that maximizes $\Gamma(\boldsymbol{\Theta}|\widetilde{\boldsymbol{\Theta}})$ with respect to $\boldsymbol{\Theta}$, that is, $\boldsymbol{\Theta}_{new} = \arg\max_{\boldsymbol{\Theta}} \Gamma(\boldsymbol{\Theta}|\widetilde{\boldsymbol{\Theta}})$. It follows from (4.3) that

$$\pi_{0,new} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T} p(Z_{i,t}=0|O_{i,t},A_{i,t-1},O_{i,t-1};\widetilde{\boldsymbol{\Theta}}_0),$$

$$\pi_{v,new} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T} p(Z_{i,t}=1|O_{i,t},v_t;\widetilde{\boldsymbol{\Theta}}_v),$$

$$\pi_{u,new} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T} p(Z_{i,t}=2|O_{i,t},u_i;\widetilde{\boldsymbol{\Theta}}_u).$$

The rest of the parameters can be updated using any derivative-based (e.g., quasi-Newton) or derivative-free (e.g., Nelder-Mead) algorithm. Our final estimator is obtained by repeating the E-step and the M-step until convergence.

Choice of the parametric family. In our implementation, when the observation is continuous, we posit normal distribution functions for $p_{u_i}(o'|u_i)$, $p_{v_t}(o'|v_t)$, $p_0(o'|a,o)$, that is, $p_{u_i}(o'|u_i) = \phi(o';\mu_{u_i},\Sigma_{u_i})$, $p_{u_i}(o'|v_t) = \phi(o';\mu_{v_t},\Sigma_{v_t})$ and $p_0(o'|a,o) = \phi(o';\mu_0(a,o),\Sigma_0(a,o))$ where $\phi(\bullet;\mu,\Sigma)$ denotes a d-dimensional multivariate normal density function with mean vector μ and covariance matrix Σ. We further use a linear model for the mean function μ_0, that is, $\mu_0(a,o) = \Lambda o + \psi a$ and a constant model for the covariance function, that is, $\Sigma_0(a,o) = \Sigma_0$ for any a and o. As such, the set of parameters $\boldsymbol{\Theta}$ can be summarized by $\{\pi_0,\pi_v,\pi_u,\{\mu_{u_i}\}_i,\{\Sigma_{u_i}\}_i,\{\mu_{v_t}\}_t,\{\Sigma_{v_t}\}_t,\Lambda,\psi,\Sigma_0\}$.

5. Theoretical Results

In this section, we focus on investigating the theoretical properties of our proposed model-free estimators. Consistencies and convergence rates of the model-based estimators can be established based on existing analyses of EM algorithms (see e.g., Wu 1983; Balakrishnan, Wainwright, and Yu 2017) and we omit the details to save space.

Summary. We begin with a summary of our theoretical results. Theorem 1 presents the convergence rates of the proposed value estimators. In particular, for a sufficiently large L, we show that the estimated average effect $\widehat{\eta}^{\pi}$, individual-specific effect $\widehat{\eta}^{\pi}_i$, time-specific effect $\widehat{\eta}^{\pi}_t$ and individual- and time-specific $\widehat{\eta}^{\pi}_{i,t}$ converge at a rate of $(NT)^{-1/2}$, $T^{-1/2}$, $N^{-1/2}$ and $\min^{-1/2}(N,T)$, respectively, up to some logarithmic factors. Theorem 2 establishes the limiting distributions of these estimators. We next impose some technical assumptions.

Assumption 1 (Realizability). Assume that there exist some constants p and C, such that for any $a\in\mathcal{A}$, the reward function $r_1(\cdot,a)\in\Lambda(p,C)$, where $\Lambda(p,C)$ is the Hölder class with the smoothness parameter p (see Section B.2.1 for the definition).

Assumption 2 (Completeness). For any function g such that $g(\cdot,a) \in \Lambda(p,C)$ for all action a, $\mathcal{B}^{\pi}g(\cdot,a) \in \Lambda(p,C)$ where $\mathcal{B}^{\pi}$ denotes the Bellman operator, that is, $(\mathcal{B}^{\pi}g)(o,a) = \sum_{a'}\mathbb{E}_{O'\sim p_0(O'|o,a)}[\pi(a'|O')g(O',a')]$.

Assumption 3 (Basis functions). (i) $\sup_o \|\boldsymbol{\Phi}_L(o)\|_2 = O(\sqrt{L})$ and $\lambda_{\max}[\int_{o\in\mathcal{O}}\boldsymbol{\Phi}_L(o)\boldsymbol{\Phi}_L^{\top}(o)do] = O(1)$; (ii) For any $C>0$, $\sup_{f\in\Lambda(p,C)}\inf_{\beta\in\mathbb{R}^L}\sup_o|\Phi_L^{\top}(o)\beta - f(o)| = O(L^{-p/d})$; (iii) $L \ll \min(N,T)/\log(NT)$; (iv) $NT \ll L^{2p/d}$.

Assumption 4 (System dynamics). (i) Assume that there exist random errors $\{e_{i,t}\}_{i,t}$ that are iid copies of E such that the future observation $O_{i,t+1}$ can be represented as $\kappa(O_{i,t},A_{i,t},u_i,v_t,e_{i,t})$ for some function κ that satisfies

$$\sup_{a,u,v}\mathbb{E}\|\kappa(o,a,u,v,E)-\kappa(o',a,u,v,E)\|_2 \le q\|o-o'\|_2,$$

$$\sup_{o,a}\|\kappa(o,a,u,v,E)-\kappa(o,a,u,v,E')\|_2 = O(\|E-E'\|_2),$$

for some $0\le q<1$; (ii) each element of the error E vector has sub-exponential tail, that is, $\max_j \mathbb{E}\exp(t|E_j|) < \infty$ for some $t>0$, where E_j denotes the jth element of E; (iii) the reward function r_1, the density functions p_0, p_u, p_v and $N^{-1}\sum_{i=1}^{N}p_{O_{i,1}}$ ($p_{O_{i,1}}$ denotes the density function of $O_{i,1}$) are uniformly bounded. (iv) there exists some constant $c\ge 1$ such that $\mathbb{E}(\sum_{k=1}^{T}\varepsilon^2_{k,i,t}|O_{i,t}=o,A_{i,t}=a) > c^{-1}$, for any $i,t,o\in\mathcal{O}$, $a\in\mathcal{A}$, where $\varepsilon_{k,i,t}$ denotes the Bellman error defined in Section B.3 of the supplementary article.

Assumption 5 (Stability). For any backward step k (the kth iteration in Algorithm 1),

$$\lambda_{\min}[\mathbb{E}(\boldsymbol{\Phi}_k^{\top}\boldsymbol{S}_k\boldsymbol{\Phi}_{k-1}^{new})] \ge (NT)\rho_0 \text{ and}$$

$$\|[\mathbb{E}(\boldsymbol{\Phi}_k^{\top}\boldsymbol{S}_k\boldsymbol{\Phi}_k)]^{-1}\mathbb{E}(\boldsymbol{\Phi}_k^{\top}\boldsymbol{S}_k\boldsymbol{\Phi}_{k-1}^{new})\|_2 \le \rho_1,$$

for some constants $\rho_0>0$ and $0<\rho_1<1$, where $\boldsymbol{\Phi}_k$ is the matrix consisting of the first $N(T-k+1)$ rows of matrix $\boldsymbol{\Phi}$, $\boldsymbol{S}_k$ and $\boldsymbol{B}_k$ are the residual maker matrix and the design matrix for the fixed effects at step k, respectively (see Section B.2.4.5 of the supplementary article for the detailed formulation), and the matrix $\boldsymbol{\Phi}_k^{new}$ is a variant of $\boldsymbol{\Phi}_k$, with its detailed definition provided in Section B.2.4.5 of the supplementary article.

The realizability and the Bellman completeness assumptions are commonly imposed in the RL literature (see e.g., Chen and Jiang 2019; Uehara, Shi, and Kallus 2022). Realizability essentially requires the Hölder class to be sufficiently rich to contain r_1. The Bellman completeness requires the Hölder class to be "complete" in the sense that it remains closed under the Bellman operator. It holds automatically when the transition function belong to the Hölder class as well. The Hölder smoothness assumption is frequently imposed in the sieve estimation literature (Huang 1998; Chen and Christensen 2015). It has seen increasing adoption in the RL literature as well (Fan et al. 2020; Chen and Qi 2022; Shi et al. 2022). Assumption 3(i) and (ii) are automatically satisfied when tensor product B-spline or wavelet bases is used; see Section 6 of Chen and Christensen (2015) for a review of these basis functions. Assumption 3(i) upper bounds the ℓ_2 norm of the sieve basis vector and the maximum eigenvalue of $\int_{o\in\mathcal{O}} \boldsymbol{\Phi}_L(o)\boldsymbol{\Phi}_L^\top(o)do$ whereas Assumption 3(ii) upper bounds the approximation error of the sieve estimator. Assumption 3(iii) upper bounds the number of basis functions and is to guarantee the consistency of the estimator. Assumption 3(iv) lower bounds the number of basis functions, requiring the bias of our estimator to be much smaller than its standard deviation so as to establish its asymptotic normality.

Assumption 4(i) requires κ to be Lipschitz continuous. Assumption 4(ii) requires the error distribution to possess the sub-exponential tail. These two conditions allow us to establish concentration inequalities in doubly inhomogeneous environments. Under the additivity assumption (2.3), there exist functions κ_0, κ_u, κ_v and random errors E_0, E_u, and E_v such that $\kappa(o,a,u,v,E) \overset{d}{=} \mathbb{I}(Z=0)\kappa_0(o,a,E_0) + \mathbb{I}(Z=1)\kappa_v(v,E_v) + \mathbb{I}(Z=2)\kappa_u(u,E_u)$ where the latent variable Z is independent of (E_0,E_u,E_v) and satisfies that $\mathbb{P}(Z=0)=\pi_0$, $\mathbb{P}(Z=1)=\pi_v$, $\mathbb{P}(Z=2)=\pi_u$. As such, Assumption 4(i) and (ii) are automatically satisfied if E_0, E_v, and E_u have sub-exponential tails, κ_0, κ_u and κ_v are Lipschitz continuous as functions of the error term and

$$\sup_a \mathbb{E}\|\kappa_0(o,a,E_0) - \kappa_0(o',a,E_0)\|_2 \le q\|o-o'\|_2, \qquad (5.1)$$

for some $0 \le q < 1$. Notice that (5.1) is automatically satisfied for the auto-regressive model $O' = f(O,A)+g(E_0)$ for any f such that $\sup_a |f(o,a)-f(o',a)| \le q\|o-o'\|_2$. Other examples are provided in Diaconis and Freedman (1999). Assumption 4(iii) requires the density of the latent factors and the initial observations to be upper bounded, thus yielding the uniform boundedness of marginal density functions of $\{O_{i,t}\}_{i,t}$. Assumption 4(iv) lower bounds the second moment of the temporal difference error to invoke the martingale central limit theorem (McLeish 1974) to establish the asymptotic normality of our estimator.

The first part of Assumption 5 essentially requires $(NT)^{-1}\mathbb{E}(\boldsymbol{\Phi}_k^\top \boldsymbol{S}_k \boldsymbol{\Phi}_k)$ to be invertible. The second part is closely related to the irrepresentable or mutual incoherence condition in the variable selection literature for the selection consistency of the least absolute shrinkage and selection operator (Meinshausen and Bühlmann 2006). It imposes a norm constraint on the regression coefficients of the irrelevant predictors $\boldsymbol{\Phi}_{k-1}^{new}$ on the relevant predictors $\boldsymbol{\Phi}_k$. This type of assumption is necessary to ensure the consistency of the subsequent value estimator (Perdomo et al. 2022). Similar assumptions have been imposed in the statistics literature (Luckett et al. 2020; Shi et al. 2022).

Results. Finally, we present our theories. Recall that both η_i^π, $\eta_{i,t}^\pi$ as well as their estimators implicitly depend on the initial observation $O_{i,1}$. As such, it is proper to write them as functions of $O_{i,1}$, for example, $\eta_{i,t}^\pi(o) = \mathbb{E}^\pi(R_{i,t}|O_{i,1}=o)$, $\widehat{\eta}_{i,t}^\pi(o) = \sum_a \widehat{Q}_{i,1,t}^\pi(o,a)\pi(a|o)$ ($\eta_i^\pi(o)$ and $\widehat{\eta}_i^\pi(o)$ can be similarly defined). For these values, instead of considering the differences $\widehat{\eta}_{i,t}^\pi(O_{i,1}) - \eta_{i,t}^\pi(O_{i,1})$ and $\widehat{\eta}_{i,t}^\pi(O_{i,1}) - \eta_i^\pi(O_{i,1})$, we focus on the aggregated differences $\int_{o\in\mathcal{O}}[\widehat{\eta}_{i,t}^\pi(o) - \eta_{i,t}^\pi(o)]do$ and $\int_{o\in\mathcal{O}}[\eta_i^\pi(o) - \widehat{\eta}_{i,t}^\pi(o)]do$ to eliminate the variability due to $O_{i,1}$.

Theorem 1 (Rates of Convergence). Assume Assumptions 2, 3(i)–(iii), 4(i)–(iii), and 5 hold. Then with probability approaching 1, we have for any $1 \le i \le N$ and $1 \le t \le T$,

$$\begin{aligned}
&\max_{i,t}\left|\int_{o\in\mathcal{O}}[\widehat{\eta}_{i,t}^\pi(o) - \eta_{i,t}^\pi(o)]do\right| \\
&\quad = O(L^{-p/d}) + O(\sqrt{\log(NT)/N}) + O(\sqrt{\log(NT)/T}), \\
&\max_{i}\left|\int_{o\in\mathcal{O}}[\widehat{\eta}_{i}^\pi(o) - \eta_{i}^\pi(o)]do\right| \\
&\quad = O(L^{-p/d}) + O(\sqrt{\log(NT)/T}), \\
&\max_t |\widehat{\eta}_t^\pi - \eta_t^\pi| = O(L^{-p/d}) + O(\sqrt{\log(NT)/N}) \text{ and} \\
&|\widehat{\eta}^\pi - \eta^\pi| = O(L^{-p/d}) + O(\sqrt{\log(NT)/NT}).
\end{aligned}$$

Theorem 1 highlights a noteworthy property of our method: the error bounds of value estimator depend solely on p, d, L, N, and T, and are independent of the number of backward inductions conducted. This is due to an important feature of our approach: the error term at the kth backward stage is of order $O(\pi_0^k)$ (as demonstrated by Lemma 1 in the supplementary article). Specifically, the error bounds for each value estimator comprise two components: the bias term $O(L^{-p/d})$ and the variance term $O(N^{-1/2}\sqrt{\log(NT)})$, $O(T^{-1/2}\sqrt{\log(NT)})$ or $O(\sqrt{\log(NT)/NT})$. The bias term quantifies the approximation error incurred by using linear sieves to approximate the underlying Q-function. Evidently, this bias term diminishes as the smoothness parameter p increases. As such, it implies that the smoother the system dynamics are, the smaller the approximation error becomes. Moreover, for sufficiently large L, it is evident that due to aggregation over time and population, the average effect $\widehat{\eta}^\pi$ converges the fastest, whereas the individual- and time-specific effect $\widehat{\eta}_{i,t}^\pi$ demonstrates a relatively slower convergence.

Theorem 2 (Asymptotically Normality). Assume Assumptions 2–5 hold. Then when both N and T goes to infinity,

$$\begin{aligned}
&\sqrt{\min(N,T)}\sigma_{\eta_{i,t}^\pi}^{-1}\int_{o\in\mathcal{O}}\left(\widehat{\eta}_{i,t} - \eta_{i,t}\right)do \xrightarrow{d} \mathcal{N}(0,1), \\
&\sqrt{T}\sigma_{\eta_i^\pi}^{-1}\int_{o\in\mathcal{O}}(\widehat{\eta}_i^\pi - \eta_i^\pi)do \xrightarrow{d} \mathcal{N}(0,1), \\
&\sqrt{N}\sigma_{\eta_t^\pi}^{-1}(\widehat{\eta}_t^\pi - \eta_t^\pi) \xrightarrow{d} \mathcal{N}(0,1), \text{ and} \\
&\sqrt{NT}\sigma_{\eta^\pi}^{-1}(\widehat{\eta}^\pi - \eta^\pi) \xrightarrow{d} \mathcal{N}(0,1),
\end{aligned}$$

where $\sigma_{\eta_{i,t}^\pi}, \sigma_{\eta_i^\pi}, \sigma_{\eta_t^\pi}$ and σ_{η^π} are some quantities bounded from below and above (for a detailed formulation, refer to Section B.3 in the supplementary article).

Theorem 2 establishes the asymptotic normality of the value estimators when both N and T diverges. It lays the foundations for statistical inference (e.g., constructing confidence intervals) of these policy values. Specifically, one could estimate the standard deviations $\sigma_{\eta_{i,t}^\pi}, \sigma_{\eta_i^\pi}, \sigma_{\eta_t^\pi}$, σ_{η^π} from the data, and then employ these estimators to construct Wald-type confidence intervals.

6. Simulation Studies

In this section, we evaluate our proposed model-based and model-free approaches through extensive simulations. We begin by specifying the competing methods. Next, we evaluate the performance of our model-free method using the RL benchmark dataset D4RL (Fu et al. 2020). Finally, we investigate the sensitivity of the proposed model-free and model-based estimators to the additivity assumption. We focus on evaluating the following four targets: (i) the average reward η^π; (ii) the ith subject's average reward aggregated over time η_i^π; (iii) the average reward in the population at time t η_t^π; (iv) the ith subject's expected reward at time t, denoted by $\eta_{i,t}^\pi$. Throughout, we evaluate these four targets for all subjects over the first five time periods in the offline dataset. Additional numerical results and details about the environments, can be found in Section C of the supplementary article.

Competing methods. We compare our proposed approaches against the following methods, including two direct methods (DM), three importance sampling (IS) methods, three doubly robust (DR) methods, and one model-based (MB) method:

(i) DM1: an adaptation of fitted Q-evaluation (Le, Voloshin, and Yue 2019) to the average reward;
(ii) DM2 : an adaptation of Q-function based least-squares temporal difference (see, e.g., Shi et al. 2022) to the average reward setting;
(iii) IS1: sequential IS that uses the product of IS ratios at each time to address the distributional shift between the behavior and target policies (Precup 2000);
(iv) IS2: marginalized IS that replaces the product of IS ratios with the marginalized IS ratio to break the curse of horizon (Liu et al. 2018; Kallus and Uehara 2020);
(v) IS3: marginalized IS based on minimax weight learning (Uehara, Huang, and Jiang 2020);
(vi) DR1: a doubly robust method that employs the influence function developed by Jiang and Li (2016) to construct the estimator and uses approaches from DM1 and IS1 to compute the Q-function and the IS ratio;
(vii) DR2: a doubly robust method that employs the influence function developed by Kallus and Uehara (2020) to construct the estimator and uses approaches from DM1 and IS2 to compute the Q-function and the IS ratio;
(viii) DR3: a doubly robust method that employs the influence function developed by Liao et al. (2022) to construct the estimator and uses approaches from DM2 and IS3 to compute the Q-function and the IS ratio;
(ix) MB: a standard model-based method developed in doubly homogeneous environments.

We also remark that the three IS and DR methods primarily differ in their utilization of IS ratios. Specifically, IS1 and DR1 use the sequential IS ratio, requiring the environment to be individually homogeneous—meaning all individuals' data trajectories share the same distribution. In contrast, IS2 and DR2 employ the marginalized IS ratio, whose validity additionally depends on the Markov assumption. Meanwhile, IS3 and DR3 apply another marginalized IS ratio that further requires the stationarity assumption. However, all the three aforementioned assumptions are violated in doubly inhomogeneous environments due to the presence of $\{U_i\}_i$ and $\{V_t\}_t$. Further details of these methods are relegated to Section C.1 of the supplementary article.

Application to D4RL. D4RL consists of a collection of benchmark datasets specifically designed for evaluating RL algorithms. Its primary goal is to provide standardized and diverse data that assist researchers and practitioners in developing advanced methodologies for offline RL. We evaluate the performance of our method across four D4RL environments: Maze2D, Hopper, HalfCheetah, and Walker2d. For each environment, we further consider four distinct settings. For Maze2D, the settings differ in maze layouts and the level of difficulty in reaching the goal state. The four specific settings we consider include "open," "umaze," "medium," and "large". For HalfCheetah, Walker2D and Hopper, the settings are defined by varying behavior policies, and we consider the four settings labeled "noisy," "medium," "medium-replay," and "medium-expert". The datasets can be directly downloaded from *http://rail.eecs.berkeley.edu/datasets/offline_rl/*. More details can be found in Section C.2 of the supplementary article and the D4RL Wiki page[2].

In all settings, the target policy we aim to evaluate is fixed to a randomized policy that follows a uniform distribution across the action space. To simulate doubly inhomogeneous environments, we inject two-way fixed effects into the original reward $R_{i,t}$ from the D4RL datasets, leading to the modified reward $\widetilde{R}_{i,t} = R_{i,t} + \cos(t) + \sin(i)$. All observations and actions from the original data remain unchanged. To ensure fair comparison, the Q-functions in the proposed, direct and doubly robust methods are all modeled via linear sieves with a quadratic basis function. For IS and DR, we use a conditional Gaussian model with a linear conditional mean function and constant variance to approximate the behavior policy.

MSEs of various model-free estimators are reported in Tables 1 and 2[3]. Recall that for each environment, we consider four settings, each containing four evaluation targets. This yields 16 cases for each environment. Overall, the proposed model-free method (denoted by P1) achieves the best performance in most cases:

- In Maze2D, our proposed method ranks first in 15 out of 16 cases;
- In HalfCheetah, our method ranks first in 14 out of 16 cases;
- In Walker2D, our proposed method ranks first in 12 out of 16 cases;
- In Hopper, our method ranks first in 8 out of 16 cases.

Meanwhile, there are a few exceptions: In the Hopper-noisy setting, DR3 outperforms our method for evaluating all the four estimands. Likewise, for Hopper-medium-replay, IS1 achieves the best performance for estimating η^π, η_i^π, and η_t^π. Despite these specific cases, our method exhibits superior performance in general. It is also worth noting that our proposed method is not overly sensitive to diverse behavior policies. In contrast,

[2] *https://github.com/Farama-Foundation/d4rl/wiki/Tasks*.

[3] To enhance clarity, we remove the two model-based methods in the tables, as their performance are much worse than the model-free methods, due to the difficulty in accurately modeling the complex transition function in D4RL.

Table 1. MSEs of the estimated value (four targets) using our proposed methods and other competing methods for Maze2D and Halfcheetah with $N = T = 20$ over 20 replications.

	Maze2D-open				Maze2D-umaze				Maze2D-medium				Maze2D-large			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
P1	0.01	0.45	0.30	0.75	0.02	0.47	0.28	0.73	0.00	0.45	0.31	0.76	0.00	0.45	0.32	0.77
DM1	0.01	0.49	0.34	0.83	0.03	0.52	0.33	0.82	0.01	0.51	0.35	0.85	0.00	0.50	0.36	0.85
DM2	3.75	4.25	3.72	4.23	2.98	3.49	3.93	4.43	0.75	1.26	1.12	1.64	0.55	1.07	0.96	1.48
IS1	0.66	1.17	1.26	3.63	0.42	0.93	0.39	2.06	0.35	0.87	0.62	2.56	0.62	1.13	1.12	3.34
IS2	1.52	2.03	6.10	10.12	1.81	2.32	4.65	8.06	0.93	1.44	3.43	6.67	1.28	1.80	5.22	8.94
IS3	0.01	0.52	0.35	0.85	0.03	0.54	0.33	0.84	0.01	0.52	0.35	0.87	0.00	0.52	0.36	0.87
DR1	0.25	2.99	0.44	7.03	0.99	12.81	3.11	60.60	0.15	1.80	0.38	7.45	0.21	1.41	0.28	4.31
DR2	0.25	3.09	1.16	13.04	0.13	2.68	0.65	9.86	0.18	2.35	0.64	8.82	0.21	1.95	0.64	8.06
DR3	0.01	0.51	0.36	0.86	0.03	0.54	0.33	0.84	0.01	0.52	0.36	0.87	0.00	0.52	0.36	0.88

	Halfcheetah-medium				Halfcheetah-noisy				Halfcheetah-medium-replay				Halfcheetah-medium-expert			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
P1	0.06	0.56	0.44	0.95	0.06	0.57	0.44	0.95	0.05	0.57	0.46	0.97	0.06	0.57	0.45	0.96
DM1	0.06	0.56	0.44	0.95	0.06	0.57	0.45	0.96	0.05	0.57	0.46	0.97	0.06	0.57	0.45	0.96
DM2	1.46	1.96	1.92	2.43	0.36	0.87	0.87	1.38	0.50	1.02	1.01	1.52	1.46	1.97	1.92	2.44
IS1	0.06	0.57	0.47	0.99	0.27	0.77	1.07	1.89	0.08	0.60	0.46	1.06	0.06	0.58	0.48	1.00
IS2	0.05	0.56	0.45	0.96	0.63	1.14	2.69	3.46	0.10	0.62	0.95	1.87	0.05	0.57	0.46	0.97
IS3	0.05	0.56	0.46	0.97	0.06	0.57	0.48	0.99	0.05	0.57	0.46	0.97	0.06	0.57	0.46	0.98
DR1	0.15	0.77	0.86	1.92	3.82	3.13	4.25	6.34	0.38	2.88	1.77	12.80	0.15	0.78	0.87	1.92
DR2	0.06	0.57	0.44	0.96	0.06	0.60	0.45	1.08	0.10	0.80	0.52	2.50	0.06	0.57	0.45	0.97
DR3	0.08	0.59	0.51	1.02	0.22	0.73	0.70	1.21	0.05	0.57	0.46	0.98	0.09	0.60	0.51	1.02

NOTE: The best method with smallest MSE in each column were highlighted with blue.

Table 2. MSEs of the estimated value (four targets) using our proposed methods with other competing methods for Walker2D and Hopper with $N = T = 20$ over 20 replications.

	Walker2D-medium				Walker2D-noisy				Walker2D-medium-replay				Walker2D-medium-expert			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
P1	0.51	1.03	0.55	1.07	0.52	1.03	0.55	1.07	0.53	1.03	0.55	1.06	0.52	1.03	0.55	1.07
DM1	0.52	1.03	0.56	1.08	0.52	1.04	0.55	1.07	0.53	1.04	0.56	1.07	0.52	1.04	0.56	1.08
DM2	38.93	39.45	38.60	39.12	22.15	22.66	23.29	23.80	31.70	32.21	39.07	39.58	38.94	39.46	38.60	39.11
IS1	0.52	1.03	0.57	1.09	8.31	8.82	51.74	57.08	0.53	1.04	0.56	1.07	0.53	1.04	0.57	1.08
IS2	0.52	1.03	0.57	1.09	0.36	0.87	0.40	1.42	0.53	1.04	0.55	1.08	0.53	1.04	0.57	1.08
IS3	0.52	1.03	0.57	1.09	0.53	1.04	0.57	1.08	0.53	1.04	0.57	1.07	0.53	1.04	0.57	1.08
DR1	0.56	1.07	1.02	1.55	3.09	6.55	22.45	28.16	0.59	1.12	0.98	1.58	0.56	1.08	1.02	1.55
DR2	0.52	1.03	0.56	1.08	0.48	1.00	0.48	1.53	0.53	1.06	0.56	1.18	0.52	1.04	0.56	1.08
DR3	0.52	1.03	0.56	1.09	0.53	1.04	0.57	1.08	0.52	1.03	0.56	1.07	0.52	1.04	0.56	1.08

	Hopper-medium				Hopper-noisy				Hopper-medium-replay				Hopper-medium-expert			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
P1	0.43	0.92	0.76	1.25	0.38	0.87	0.67	1.17	0.43	0.92	0.76	1.25	0.42	0.90	0.76	1.25
DM1	0.44	0.96	0.79	1.30	0.39	0.90	0.69	1.21	0.44	0.96	0.80	1.31	0.44	0.94	0.80	1.31
DM2	40.91	41.44	40.98	41.50	4.02	4.53	4.29	4.81	26.61	27.13	27.09	27.60	21.99	22.49	22.29	22.80
IS1	0.48	1.00	0.86	1.39	0.48	1.00	2.12	2.85	0.37	0.89	0.70	1.42	0.55	1.05	1.12	1.93
IS2	0.48	1.00	0.87	1.39	0.42	0.93	1.17	1.88	0.41	0.93	0.78	1.58	0.46	0.96	0.82	1.61
IS3	0.46	0.98	0.86	1.37	0.44	0.95	0.83	1.35	0.46	0.98	0.86	1.37	0.45	0.96	0.85	1.36
DR1	0.45	0.97	0.81	1.32	0.41	1.23	0.75	2.54	0.50	1.11	0.82	1.91	0.39	1.47	0.81	4.77
DR2	0.44	0.95	0.79	1.30	0.59	1.39	1.05	3.18	0.46	1.04	0.81	2.33	0.51	1.18	0.84	2.43
DR3	3.48	4.00	3.88	4.39	0.22	0.74	0.60	1.12	0.93	1.45	1.30	1.81	8.42	8.92	8.75	9.26

NOTE: The best method with the smallest MSE in each column is highlighted in blue.

competing methods like IS1 and IS2, which require to learn the behavior policy, vary considerably across different settings.

Sensitivity analysis. We have designed four synthetic environments with binary actions and continuous observations to investigate the sensitivity of our model-free (P1) and model-based (P2) estimators to the additivity assumption. These environments differ from D4RL in that, unlike the D4RL where only the reward is doubly inhomogeneous, in these environments, we introduce latent factors into the transition function to make it doubly inhomogeneous as well.

Specifically, we consider two reward models: an additive model (2.4) and a factor model (3.5). In the additive model, the two-way fixed effects θ_i and λ_t are set to $\sin(i)$ and $\cos(t)$, respectively. In the factor model, we set $\boldsymbol{\gamma}_i = (\sin(i), \sin(2i), \sin(3i))^\top$ and $\boldsymbol{\alpha}_t = (\cos(t), \cos(2t), \cos(3t))^\top$. In both models, the reward function is fixed to $r_1(o,a) = a - 0.25o$ and residuals $\varepsilon_{i,t}$s are iid Gaussian random errors with mean zero and variance 0.25.

Moreover, we consider three transition models: (i) an additive model where $O_{i,t+1} = -0.25O_{i,t} + A_{i,t} + \sin(i) + \cos(t) + e_{i,t}$; (ii) a factor model where $O_{i,t+1} = -0.25O_{i,t} + A_{i,t} + \boldsymbol{\gamma}_i^\top \boldsymbol{\alpha}_t + e_{i,t}$;

(iii) a regime switching model where

$$O_{i,t+1} = \begin{cases} -0.25O_{i,t} + A_{i,t} + 2e_{i,t}, & \text{if both } i \text{ and } t \text{ is odd,} \\ O_{i,t} - A_{i,t} + e_{i,t}, & \text{if } i \text{ is odd, and } t \text{ is even,} \\ 0.25O_{i,t} - A_{i,t} + 2e_{i,t}, & \text{if } i \text{ is even, and } t \text{ is odd,} \\ -O_{i,t} + A_{i,t} + e_{i,t}, & \text{otherwise.} \end{cases}$$

Similarly, $e_{i,t}$s are set to iid Gaussian errors with mean zero and variance 0.25.

Table 3 summarizes reward and transition models of the four environments. It can be seen that in each environment, either the reward or the transition model does not satisfy the additive structure, leading to the violation of the proposed model assumption. For all the environments, we set $N = 40$, $T = 40$. The behavior policy is a uniform random policy whereas the target policy is another random policy where $\pi(1|o) = 0.8$ for any o. The results are reported in Table 4. In first three environments where the additivity assumption holds for either the reward or the transition model, our proposed method generally outperforms the competing methods in estimating η_i^π, η_t^π, and $\eta_{i,t}^\pi$. In the last environment where both models are interactive, our proposed method no longer dominates other methods, but its performance remains comparable to those of the best competing methods.

7. Real Data Analysis

In this section, we apply our proposed method to a sepsis dataset from MIMIC-III (Johnson et al. 2016), a database that contains information on critical care patients from Beth Israel Deaconess Medical Center in Boston, MA. As mentioned earlier, the heterogeneity in patients' response to treatment (Evans et al. 2021), along with potentially non-stationary environments makes it difficult to consistently assess the impact of conducting a given target policy on patient outcomes.

Table 3. A summary of environments in the sensitivity analysis.

Environment	I	II	III	IV
Reward	Additive	Additive	Factor	Factor
Transition	Regime switching	Factor	Additive	Factor

We focus on a subset of patients who received treatments sequentially over 20 stages. The primary outcome in this analysis is the sequential organ failure assessment (SOFA) score (Jones, Trzeciak, and Kline 2009), which monitors the progression of organ failure over time and measures the degree of organ dysfunction or failure in critically ill patients. A higher SOFA score indicates a higher risk of mortality. At any time point t, we consider a binary treatment $A_t \in \{0, 1\}$ where $A_t = 1$ indicates that the patient received an intravenous fluid intervention with a dose greater than the median value for the group of patients being studied, and $A_t = 0$ otherwise. In previous studies, Zhou et al. (2022) examined joint action spaces with both vasopressors and intravenous fluid interventions. We focus solely on the intravenous fluid intervention in light of the findings of Zhou et al. (2022), which detected a limited impact of vasopressors.

The following five covariates are included in the analysis: gender, age, the Elixhauser comorbidity index, weight, and the systemic inflammatory response syndrome score. Three deterministic policies were evaluated using our proposed methods: (i) always administering a high dose, (ii) always administering a low dose, and (iii) administering a low dose when the SOFA score is less than 11, and a high dose otherwise. The third policy is tailored to the SOFA score, taking into account evidence that a SOFA score of more than 11 is associated with a 100% mortality rate (Jones, Trzeciak, and Kline 2009). To estimate the Q-function, we employed a second-order degree polynomial two-way fixed effects model at each iteration. The average value estimators for the three policies are as follows: 7.26 (always high dose), 6.85 (always low dose), and 6.51 (tailored by SOFA score). These results indicate that the tailored policy is the most effective policy as it yields the lowest estimated SOFA score. Figure 1 summarizes the estimated η_i^πs and η_t^πs, clearly demonstrating that the tailored policy outperforms the other two policies, while the always high dose policy performs the worst. Our conclusion is in line with these existing results, which recommend the low dose policy over the high dose policy. It is also consistent with physicians' recommendations in the behavior data (Zhou et al. 2022).

Table 4. MSEs of the estimated value (four targets) using our proposed methods with other competing methods.

	Scenario 1				Scenario 2				Scenario 3				Scenario 4			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
P1	0.01	0.48	0.17	3.54	0.66	0.73	0.10	1.78	0.41	0.51	0.07	4.65	0.03	0.17	0.05	9.00
P2	0.04	0.76	0.10	3.43	0.33	1.04	0.35	2.04	0.47	0.60	0.10	4.70	0.09	0.28	0.11	9.11
MB	0.01	1.56	0.84	4.80	0.14	1.72	2.17	4.52	0.09	0.64	0.31	4.82	0.01	0.15	0.11	9.12
DM1	0.01	1.40	0.85	4.02	0.01	1.26	1.26	3.06	0.04	0.51	0.37	4.36	0.02	0.02	0.08	8.37
DM2	0.37	1.77	0.98	4.16	0.39	1.64	0.85	2.31	0.81	1.27	0.51	4.08	0.77	0.77	0.73	8.25
IS1	0.20	1.60	0.58	5.25	0.84	2.08	0.55	3.41	0.63	1.09	0.63	4.74	0.30	0.30	0.78	8.83
IS2	0.05	1.45	0.13	4.82	1.26	2.50	0.31	3.43	0.93	1.40	0.33	4.48	0.41	0.41	0.36	8.23
IS3	2.89	4.28	3.14	6.32	7.04	8.29	4.63	6.09	7.47	7.94	5.17	8.74	6.48	6.49	5.72	13.24
DR1	0.16	1.56	0.35	5.14	0.53	1.78	0.29	3.00	0.67	1.14	0.33	4.54	0.24	0.24	0.37	8.36
DR2	0.23	1.63	0.85	6.32	0.58	1.82	0.25	3.67	0.95	1.41	0.60	5.07	0.22	0.22	0.37	8.38
DR3	2.21	3.61	2.54	5.72	7.02	8.26	4.61	6.08	6.84	7.31	4.66	8.24	5.54	5.54	4.86	12.38

NOTE: The best method with the smallest MSE in each column is highlighted in blue. P1 and P2 are our proposed model-free and model-based methods, respectively.

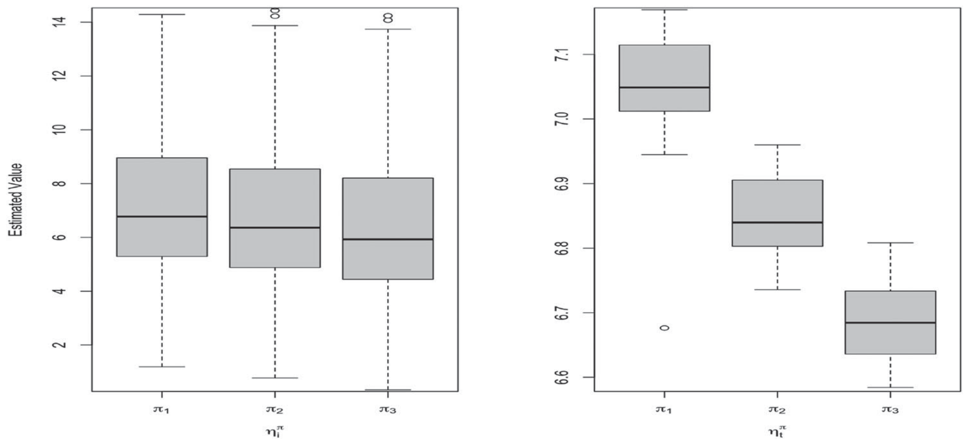

Figure 1. The estimated value function of η_i^π and η_t^π under three target policies, where π_1 is the always high dose policy, π_2 is always low dose policy, and π_3 is the tailored policy.

Supplementary Materials

The supplementary article includes the following content: Section A.1 presents a diagnostic procedure for assessing the additivity assumption. Section A.2 discusses extending our approach to evaluate history-dependent and subject-specific target policies. Section A.3 provides a comparison between model-based and model-free approaches. Sections B.1, B.2, and B.3 contain the proofs of Proposition 1, Theorem 1, and Theorem 2, respectively. Section C provides details on the baseline methods, the D4RL dataset, additional numerical results, and real data analysis.

Acknowledgments

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

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Shi's research is partially supported by an EPSRC grant EP/W014971/1. Wang's research is partly supported by an NSF grant NSF FRGMS-1952373.

References

Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press. [2]

Arkhangelsky, D., Imbens, G. W., Lei, L., and Luo, X. (2021), "Double-Robust Two-Way-Fixed-Effects Regression for Panel Data," arXiv preprint arXiv:2107.13737. [2]

Bachhuber, M. A., Saloner, B., Cunningham, C. O., and Barry, C. L. (2014), "Medical Cannabis Laws and Opioid Analgesic Overdose Mortality in the United States, 1999-2010," *JAMA Internal Medicine*, 174, 1668–1673. [3]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [3,6]

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017), "Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis," *The Annals of Statistics*, 45, 77–120. [7]

Bollen, K. A. (2002), "Latent Variables in Psychology and the Social Sciences," *Annual review of psychology*, 53, 605–634. [3]

Chen, E. Y., Song, R., and Jordan, M. I. (2022), "Reinforcement Learning with Heterogeneous Data: Estimation and Inference," arXiv preprint arXiv:2202.00088. [2]

Chen, J., and Jiang, N. (2019), "Information-Theoretic Considerations in Batch Reinforcement Learning," in *International Conference on Machine Learning*, pp. 1042–1051, PMLR. [7]

Chen, X., and Christensen, T. M. (2015), "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188, 447–465. [5,7,8]

Chen, X., and Qi, Z. (2022), "On Well-Posedness and Minimax Optimal Rates of Nonparametric q-function Estimation in Off-Policy Evaluation," in *International Conference on Machine Learning*, pp. 3558–3582, PMLR. [7]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–22. [6]

Diaconis, P., and Freedman, D. (1999), "Iterated Random Functions," *SIAM Review*, 41, 45–76. [8]

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022), "Counterfactual Inference for Sequential Experiments," arXiv preprint arXiv:2202.06891. [3]

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., French, C., Machado, F. R., Mcintyre, L., Ostermann, M., Prescott, H. C., et al. (2021), "Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021," *Intensive Care Medicine*, 47, 1181–1247. [2,11]

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020), "A Theoretical Analysis of Deep q-Learning," in *Learning for Dynamics and Control*, pp. 486–489, PMLR. [2,7]

Frisch, R., and Waugh, F. V. (1933), "Partial Time Regressions as Compared with Individual Trends," *Econometrica: Journal of the Econometric Society*, 1, 387–401. [5]

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020), "D4rl: Datasets for Deep Data-Driven Reinforcement Learning," arXiv preprint arXiv:2004.07219. [9]

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. (2019), "Combining Parametric and Nonparametric Models for Off-Policy Evaluation," in *International Conference on Machine Learning*, pp. 2366–2375, PMLR. [1]

Green, D. P., Kim, S. Y., and Yoon, D. H. (2001), "Dirty Pool," *International Organization*, 55, 441–468. [3]

Hotz, V. J., and Xiao, M. (2011), "The Impact of Regulations on the Supply and Quality of Care in Child Care Markets," *American Economic Review*, 101, 1775–1805. [3]

Hu, L., Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. (2022), "Doubly Inhomogeneous Reinforcement Learning," arXiv preprint arXiv:2211.03983. [1,2]

Huang, J. Z. (1998), "Projection Estimation in Multiple Regression with Application to Functional Anova Models," *The Annals of Statistics*, 26, 242–272. [5,7]

Imai, K., and Kim, I. S. (2019), "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science*, 63, 467–490. [2]

——— (2021), "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 29, 405–415. [2,3]

Jiang, N., and Li, L. (2016), "Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning," in *International Conference on Machine Learning*, pp. 652–661, PMLR. [1,9]

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., et al. (2016), "Mimic-iii, A Freely Accessible Critical Care Database," *Scientific Data*, 3, 1–9. [1,11]

Jones, A. E., Trzeciak, S., and Kline, J. A. (2009), "The Sequential Organ Failure Assessment Score for Predicting Outcome in Patients with Severe Sepsis and Evidence of Hypoperfusion at the Time of Emergency Department Presentation," *Critical Care Medicine*, 37, 1649–1654. [11]

Kallus, N., and Uehara, M. (2020), "Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes," *Journal of Machine Learning Research*, 21, 1–63. [9]

Kallus, N., and Uehara, M. (2022), "Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning," *Operations Research*, 70, 3282–3302. [1]

Kosorok, M. R., and Moodie, E. E. M. (2015), *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*," Philadelphia, PA: Society for Industrial and Applied Mathematics. [1]

Le, H., Voloshin, C., and Yue, Y. (2019), "Batch Policy Learning Under Constraints," in *International Conference on Machine Learning*, pp. 3703–3712, PMLR. [9]

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020), "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems," arXiv preprint arXiv:2005.01643. [1]

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023), "Is q-Learning Minimax Optimal? A Tight Sample Complexity Analysis," *Operations Research*, 72, 222–236. [2]

Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. (2022), "Reinforcement Learning in Possibly Nonstationary Environments," arXiv preprint arXiv:2203.01707. [2]

Liao, P., Klasnja, P., and Murphy, S. (2021), "Off-Policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health," *Journal of the American Statistical Association*, 116, 382–391. [1]

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. (2022), "Batch Policy Learning in Average Reward Markov Decision Processes," *The Annals of Statistics*, 50, 3364–3387. [1,9]

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* (Vol. 31). [1,9]

Lovell, M. C. (1963), "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58, 993–1010. [5]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020), "Estimating Dynamic Treatment Regimes in Mobile Health Using v-learning," *Journal of the American Statistical Association*, 115, 692–706. [1,4,8]

McLeish, D. L. (1974), "Dependent Central Limit Theorems and Invariance Principles," *The Annals of Probability*, 2, 620–628. [8]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [8]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [1,4]

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. (2022), "A Sharp Characterization of Linear Estimators for Offline Policy Evaluation," arXiv preprint arXiv:2203.04236. [8]

Precup, D. (2000), "Eligibility Traces for Off-Policy Policy Evaluation," *Computer Science Department Faculty Publication Series*, p. 80. [1,9]

Puterman, M. L. (2014), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Hoboken, NJ: Wiley. [3]

Qi, Z., Liu, D., Fu, H., and Liu, Y. (2020), "Multi-Armed Angle-based Direct Learning for Estimating Optimal Individualized Treatment Rules with Various Outcomes," *Journal of the American Statistical Association*, 115, 678–691. [1]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *Annals of Statistics*, 39, 1180–1210. [1]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, eds. D. Y. Lin and P. Heagerty, pp. 189–326, New York: Springer. [1,4]

Ross, S. (1976), "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360. [3]

Shi, C., Fan, A., Song, R., and Lu, W. (2018), "High-Dimensional A-Learning for Optimal Dynamic Treatment Regimes," *Annals of statistics*, 46, 925–957. [1]

Shi, C., Zhang, S., Lu, W., and Song, R. (2022), "Statistical Inference of the Value Function for Reinforcement Learning in Infinite-Horizon Settings," *Journal of the Royal Statistical Society*, Series B, 84, 765–793. [1,4,7,8,9]

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016), "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, 529, 484–489. [1]

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016), "The Third International Consensus Definitions for Sepsis and Septic Shock (sepsis-3)," *JAMA*, 315, 801–810. [1]

Sutton, R. S., and Barto, A. G. (2018), *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press. [1,3]

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019), *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*, Boca Raton, FL: CRC Press. [1]

Uehara, M., Huang, J., and Jiang, N. (2020), "Minimax Weight and q-function Learning for Off-Policy Evaluation," in *International Conference on Machine Learning*, pp. 9659–9668, PMLR. [1,9]

Uehara, M., Shi, C., and Kallus, N. (2022), "A Review of Off-Policy Evaluation in Reinforcement Learning," arXiv preprint arXiv:2212.06355. [1,7]

Wang, J., Qi, Z., and Wong, R. K. (2021), "Projected State-Action Balancing Weights for Offline Reinforcement Learning," arXiv preprint arXiv:2109.04640. [1]

Wang, J., Shi, C., and Wu, Z. (2023), "A Robust Test for the Stationarity Assumption in Sequential Decision Making," in *International Conference on Machine Learning*, pp. 36355–36379, PMLR. [2]

Wu, C. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [7]

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., and Ye, J. (2018), "Large-Scale Order Dispatch in On-Demand Ride-Hailing Platforms: A Learning and Planning Approach," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 905–913. [1]

Zhang, M. R., Paine, T. L., Nachum, O., Paduraru, C., Tucker, G., Wang, Z., and Norouzi, M. (2020), "Autoregressive Dynamics Models for Offline Policy Evaluation and Optimization," in *International Conference on Learning Representations*. [1]

Zhou, W., Zhu, R., and Qu, A. (2024), "Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pt-learning," *Journal of the American Statistical Association*, 119, 625–638. [1]

Zhou, Y., Qi, Z., Shi, C., and Li, L. (2022), "Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach," arXiv preprint arXiv:2210.14420. [11]