

Statistica Sinica Preprint No: SS-2026-0176	
Title	Off-Policy Evaluation with Irregularly-spaced, Outcome-dependent Observation Times
Manuscript ID	SS-2026-0176
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0176
Complete List of Authors	Xin Chen, Wenbin Lu, Shu Yang and Dipankar Bandyopadhyay
Corresponding Authors	Wenbin Lu
E-mails	wlu4@ncsu.edu
Notice: Accepted author version.	

OFF-POLICY EVALUATION WITH IRREGULARLY-SPACED, OUTCOME-DEPENDENT OBSERVATION TIMES

Xin Chen¹, Wenbin Lu², Shu Yang², and Dipankar Bandyopadhyay³

¹*Shanghai Lixin University of Accounting and Finance*

²*North Carolina State University,* ³*Virginia Commonwealth University*

Abstract: While classic off-policy evaluation literature typically assumes evenly spaced decision times for analytical convenience, this assumption is often violated in real-world applications, such as those involving user-initiated visits. To accommodate irregularly-spaced and potentially outcome-dependent decision times in off-policy evaluation, we introduce a policy evaluation criterion defined through integrated rewards rather than the traditional cumulative-reward formulation. To ensure the estimability of the value function of a policy, we further develop a new reinforcement learning framework that incorporates not only the state-action process but also an observation process, where the gap times between observations can interact with states, actions, and rewards in a more flexible way. Statistical inference for the proposed policy value is established by positing a modulated renewal process model for the observation times. The validity of the proposed approach is further supported by theoretical analysis, simulation studies, and a real-world application that uses electronic health records to evaluate periodontal

Corresponding author: Wenbin Lu. E-mail: wlu4@ncsu.edu. ORCID:

<https://orcid.org/0000-0002-7320-4755>

disease treatments.

Key words and phrases: Modulated renewal process, Policy evaluation, Reinforcement learning, Statistical inference.

1. Introduction

A dynamic policy (Azar et al., 2012) is a sequence of decision rules that guide an agent to select time-varying actions based on updated information at each decision stage throughout its interaction with the environment. Traditionally, research on dynamic policies focuses mainly on finite-horizon settings with a limited number of decision stages (Murphy, 2003; Robins, 2004), where the potential outcome framework (Rubin, 2005) is commonly used to describe some target of interest, and classical causal inference assumptions, such as unconfoundedness and consistency, are generally sufficient to guarantee the estimability of the target, provided the number of trajectories is sufficiently large. In recent years, researchers have increasingly directed their attention to dynamic policies with a relatively large number of decision stages, i.e., long- or infinite-horizon settings, as these are becoming more common in practical applications such as long-haul treatments for chronic diseases, intensive interventions during stays in the ICU, and other sequential decision-making tasks in fields like mobile health, gam-

ing, robotics, and ride-sharing. For such policies, the reinforcement learning framework (Maei and Sutton, 2010; Sutton and Barto, 2018) designed for long trajectories is preferred in tasks, such as learning the optimal policy (Luckett et al., 2019; Shi et al., 2022) and evaluating a target policy using trajectories collected from other policies, i.e., off-policy evaluation (Jiang and Li, 2016; Liu et al., 2018; Kallus and Uehara, 2020; Shi et al., 2022; Uehara et al., 2022). In particular, by assuming a Markov decision process (MDP) model, the value function of a dynamic policy, which is typically defined as the expected cumulative sum of discounted rewards over an infinite trajectory, can be estimated in infinite-horizon settings.

While most existing off-policy evaluation methods assume that decisions occur at fixed and regular time points, real-world scenarios often involve user-initiated visits, generating observation times that are possibly irregularly-spaced and outcome-dependent (Yang et al., 2019; Yang, 2021). For example, in our motivating periodontal disease treatment data (Clarkson et al., 2020), although patients are generally recommended to have a regular check-up every 6 months, the actual gap times between dental visits vary from 3 to 24 months. The length of the actual recall interval is correlated with the actions taken at the previous visit (such as the recommended recall interval made by dentists), as well as the outcomes

(such as the patient’s periodontal status after treatment). Similar sequential decision problems with irregularly-spaced, outcome-dependent observation times can also be found in other applications, such as warfarin dose management in stroke prevention (Fatemi et al., 2022), order dispatching in ride-sharing (Tang et al., 2019), and stochastic resource collection tasks (Du et al., 2020). Considering the underlying complex relationships between gap times, states, actions, and rewards, it is crucial to account for irregularly-spaced, outcome-dependent observation times in off-policy evaluation.

In this paper, we propose a framework for off-policy evaluation that can accommodate irregularly-spaced, outcome-dependent observation times. First, most existing off-policy evaluation methods consider the expected cumulative sum of discounted rewards as the value function for policy evaluation. This is natural when decisions are made at fixed and regular time points. However, since the magnitude of the cumulative sum of discounted rewards is often highly correlated with the frequency of observation times and the gap times between decisions, it may not provide a fair evaluation in settings with irregular observation times. To tackle this problem, we assume that the observed rewards are from a continuous reward process and consider a value function based on the integration of this continuous reward process, which

does not depend on the observation process. Similar integrated-reward-based value functions have been widely used in continuous-time decision-making settings, such as adaptive control of queueing systems (Bradtke and Duff, 1994), and preventive maintenance scheduling in production-inventory systems (Das et al., 1999).

Second, existing off-policy evaluation methods that allow for irregular observation times generally rely on the semi-Markov decision process (SMDP) framework (Bradtke and Duff, 1994; Parr, 1998), which extends the MDP framework by allowing the gap times between decisions to be stochastic and potentially irregular. For example, Du et al. (2020) used neural ordinary differential equations to construct a reinforcement learning framework for semi-Markov decision processes. Fatemi et al. (2022) applied the semi-Markov decision process framework to study offline reinforcement learning with stochastic observation times and developed its associated policy learning algorithms. However, to the best of our knowledge, none of the existing studies has explored off-policy evaluation with irregularly-spaced, outcome-dependent observation times; in particular, none have formally studied the conditions on the state-action and observation processes needed for value function estimation. To fill this gap, we construct a new framework incorporating the observation process, and establish conditions for the

state-action process and observation process aimed at ensuring the feasibility of value function estimation. The proposed framework is more flexible than the semi-Markov decision process framework, as it allows the gap times between decisions to depend on previous states, actions, and gap times, as well as to affect future states and actions. This flexibility enables us to make better use of information hidden in the observation process, and to construct more comprehensive evaluations for dynamic policies with irregularly-spaced, outcome-dependent observation times.

Finally, to estimate the proposed value function and derive its associated inference, we introduce a modulated renewal process model (Cox, 1972; Oakes and Cui, 1994) for the observation process, which characterizes the dependence structure between states, actions, and gap times. This model helps us formalize the connection between value functions based on cumulative and integrated rewards, and enables us to develop a distribution-modulated Bellman equation. Statistical inference, simulation studies, and an application to the periodontal disease treatment dataset derived from electronic health records are also provided to demonstrate the validity of the proposed off-policy evaluation method.

2. Framework

2.1 Value function

Consider a study where treatment decisions are made at irregularly spaced time points $T_0 = 0, T_1, T_2, \dots, T_k, \dots$. For each $k \geq 1$, let $X_k = T_k - T_{k-1}$ denote the gap time. For initialization, we select T_0 as a decision time for which the preceding decision time is also observed, and define X_0 as the gap time between T_0 and this preceding decision time, so that X_0 is available at T_0 . Let $S_k \in \mathbb{S}$ be the state variable that summarizes the subject's information collected up to and including time T_k . At each $T_k (k \geq 0)$, investigators can observe the status S_k and the gap time X_k , and then take an action $A_k \in \mathbb{A}$ based on S_k and X_k . The reward of action A_k is observed at T_{k+1} , and thus denoted as $R(T_{k+1})$. For simplicity, we assume that the state space $\mathbb{S}$ is a subspace of $\mathbb{R}^d$ where d is the dimension of the state vector, and the action space $\mathbb{A}$ is a discrete space $\{0, 1, \dots, m-1\}$ with m optional actions. Let $\pi(a|s, x)$ denote a policy that maps the state and gap time $(s, x) \in \mathbb{S} \times [0, \infty)$ to a probability mass function on $\mathbb{A}$, and satisfies $\pi(a|s, x) \geq 0$ and $\sum_{a' \in \mathbb{A}} \pi(a'|s, x) = 1$ for any $a \in \mathbb{A}$, $s \in \mathbb{S}$ and $x \geq 0$. Under such a policy π , a decision maker will take action $A_k = a$ with probability $\pi(a|S_k, X_k)$ for $a \in \mathbb{A}$ at T_k .

2.1 Value function

To introduce the value functions for π under irregular observations, we first present a cumulative-reward-based value function that is typically adopted in the existing reinforcement learning literature:

$$V_{\mathcal{C}}^{\pi}(s, x) \equiv E^{\pi} \left\{ \sum_{k \geq 1} \gamma^{T_k} R(T_k) \middle| S_0 = s, X_0 = x \right\}.$$

Here, $0 < \gamma < 1$ is a discount factor that ensures the existence of infinite sums and balances the immediate and long-term effects of the actions, E^{π} denotes the expectation when all the actions $\{A_k\}_{k \geq 0}$ follow policy π , and $V_{\mathcal{C}}^{\pi}(s, x)$ denotes the expected discounted cumulative reward if the state and gap time take values $S_0 = s$ and $X_0 = x$ at the initiation point $T_0 = 0$. By treating $\{T_k\}_{k \geq 0}$ as jump points of a counting process $N(t)$ and using the notation in the counting process theory, $V_{\mathcal{C}}^{\pi}(s, x)$ can be rewritten as $E^{\pi} \{ \int_0^{\infty} \gamma^t R(t) \, dN(t) | S_0 = s, X_0 = x \}$, where $R(t)$, an extension of the aforementioned notation $R(T_k)$, is redefined as a bounded reward process only observed at $\{T_k\}_{k \geq 1}$, but is potentially available for all $t \geq 0$. From this expression, we can see that the accumulation of rewards is driven by the increments of the observation process $dN(t)$. Since $E\{dN(t) | S_0 = s, X_0 = x\}$ reflects the intensity of the observation process, this representation directly reveals that the cumulative-reward-based value function depends on the observation frequency. Consequently, the cumulative-reward-based value function reflects not only the underlying reward trajectory but also char-

2.1 Value function

acteristics of the observation process. For example, if two policies induce similar reward trajectories, their values should ideally be similar regardless of the observation process. However, due to the integration with respect to $dN(t)$, the cumulative-reward-based value function may assign a larger value to a policy with more frequent observations than to one with fewer observations. Moreover, even when the two policies have the same average observation frequency, their cumulative-reward-based values may still differ if their observation times are distributed differently over the follow-up. Due to discounting, a policy with more observations early in the follow-up and fewer observations later may receive a larger value than one with the opposite observation pattern. Therefore, the cumulative-reward-based value function may not provide a proper evaluation for dynamic policies with varying observation frequencies or irregularly-spaced observation times.

To address this issue, we introduce a value function that is based on an integration of the reward process $R(t)$ into t :

$$V_{\mathcal{I}}^{\pi}(s, x) = E^{\pi} \left\{ \int_0^{\infty} \gamma^t R(t) dt \middle| S_0 = s, X_0 = x \right\}.$$

Here, by replacing $dN(t)$ with dt , this integrated-reward-based value function $V_{\mathcal{I}}^{\pi}(s, x)$ no longer directly depends on the observation process $N(\cdot)$, and thus is better suited to evaluate policies with irregularly-spaced observation times or varying observation frequencies. On the other hand,

2.1 Value function

although replacing $dN(t)$ with dt in the value function removes its direct dependence on intervention frequency or timing, the interventions can still affect the policy value through their impact on the reward trajectory $R(t)$.

Taking our motivating study of periodontal disease treatment as an example, when the outcome of interest is probing pocket depth (PPD) of a patient after treatment, the reward function $R(t)$ can be defined as the measurement of PPD at time t , and the integrated-reward-based value function $V_{\mathcal{I}}^{\pi}$ measures the patient's PPD status throughout the treatment period $[0, \infty)$.

If more frequent treatments improve a patient's oral health and lead to a better reward trajectory $R(t)$, then policies with different treatment frequencies will have different integrated-reward-based values. In contrast, if different treatment frequencies lead to similar reward trajectories, the policies will have similar integrated-reward-based values despite the differences in their observation times. Thus, $V_{\mathcal{I}}^{\pi}$ provides a more reasonable assessment of a dentist's treatment policies than $V_{\mathcal{C}}^{\pi}$. In the following sections, we will focus on developing off-policy evaluation methods based on the integrated-reward-based value function $V_{\mathcal{I}}^{\pi}$.

Lastly, we introduce two other functions closely related to $V_{\mathcal{I}}^{\pi}(s, x)$. One is the value function under a reference distribution. Let $\mathcal{G}(s, x)$ denote the reference distribution of the initiation state and gap time $\{S_0, X_0\}$.

2.2 Assumptions

Then an integrated-reward-based value function for policy π under $\mathcal{G}$ can be defined as $V_{\mathcal{I}}^{\pi}(\mathcal{G}) = \int V_{\mathcal{I}}^{\pi}(s, x) d\mathcal{G}(s, x)$. Another function that usually plays an important role in policy evaluation is the action-value function (commonly known as the Q-function). For $k \geq 0$ and $t \geq 0$, define

$$Q_{\mathcal{I},k}^{\pi}(s, x, a, t) = E^{\pi} \left\{ \int_t^{\infty} \gamma^{u-t} R(u) du \middle| S_k = s, X_k = x, A_k = a, T_k = t \right\},$$

which is the conditional expectation of the integrated reward over $[t, \infty)$, if taking treatment a at $T_k = t$, $S_k = s$ and $X_k = x$, and following the policy π afterward. It is not hard to see that $V_{\mathcal{I}}^{\pi}(s, x)$ and $V_{\mathcal{I}}^{\pi}(\mathcal{G})$ can be computed from $Q_{\mathcal{I},k}^{\pi}(s, x, a, t)$ by $V_{\mathcal{I}}^{\pi}(s, x) = \sum_{a \in \mathbb{A}} Q_{\mathcal{I},0}^{\pi}(s, x, a, 0) \pi(a|s, x)$ and $V_{\mathcal{I}}^{\pi}(\mathcal{G}) = \int \sum_{a \in \mathbb{A}} Q_{\mathcal{I},0}^{\pi}(s, x, a, 0) \pi(a|s, x) d\mathcal{G}(s, x)$.

2.2 Assumptions

To ensure the estimability of the functions $V_{\mathcal{C}}^{\pi}$, $V_{\mathcal{I}}^{\pi}$ and $Q_{\mathcal{I},k}^{\pi}$, we posit the following assumptions on the states, actions, rewards, and gap times.

(A1) For all $k \geq 0$, the next state variable S_{k+1} is generated as

$$pr(S_{k+1} \in \mathcal{B}_s | \{S_j, X_j, A_j, T_j\}_{0 \leq j \leq k}) = \mathcal{P}_S(\mathcal{B}_s; S_k, X_k, A_k)$$

where $\mathcal{B}_s \subset \mathbb{S}$ and $\mathcal{P}_S$ denotes the transition function of the next state conditional on the current state, action, and gap time.

2.2 Assumptions

(A2) For all $k \geq 0$, the next gap time X_{k+1} satisfies

$$pr(X_{k+1} \leq x | \{S_j, X_j, A_j, T_j\}_{0 \leq j \leq k}, S_{k+1}) = \mathcal{P}_X(x; S_{k+1}, S_k, X_k, A_k)$$

where $x \in \mathbb{R}^+$, and the cumulative distribution function $\mathcal{P}_X$ is absolutely continuous with respect to the Lebesgue measure.

(A3) (Conditional mean independence) For all $k \geq 0$ and for some bounded function r , the reward $R(T_{k+1})$ satisfies

$$E[R(T_{k+1}) | \{R(T_j), S_j, X_j, A_j\}_{j \leq k}, S_{k+1}, X_{k+1}] = r(S_k, X_k, A_k, S_{k+1}, X_{k+1}).$$

Assumption (A1) is similar to the classic Markov and time-homogeneous assumption for state variables, except that here the next state S_{k+1} can also depend on the previous gap time X_k . Assumption (A2) posits the Markov and time-homogeneous properties for gap times, while allowing the dependence of X_{k+1} on S_{k+1} after conditioning on S_k, X_k, A_k . For assumption (A3), the conventional conditional mean independence assumption in the MDP is $E[R(T_{k+1}) | \{R(T_j), S_j, X_j, A_j\}_{j \leq k}] = E\{R(T_{k+1}) | S_k, X_k, A_k\} = r'(S_k, X_k, A_k)$, which holds under (A3) with $r'(s, x, a) = \iint r(s, x, a, s', x') d\mathcal{P}_X(x'; s, x, a) d\mathcal{P}_S(s'; s, x, a)$. In fact, assumption (A3) is strictly stronger than the conventional conditional mean independence assumption. For example, consider a new reward $R^*(T_{k+1}) = r(S_k, X_k, A_k, S_{k+1}, X_{k+1}) +$

2.3 Comparisons with existing frameworks

$\{X_{k+1} - E(X_{k+1}|S_k, X_k, A_k)\}T_k$. Then, under assumptions (A1) and (A2), it can be shown that $E[R^*(T_{k+1})|\{R^*(T_j), S_j, X_j, A_j\}_{j \leq k}] = r'(S_k, X_k, A_k)$. However, assumption (A3) no longer holds, because the corresponding conditional mean $E[R^*(T_{k+1})|\{R^*(T_j), S_j, X_j, A_j\}_{0 \leq j \leq k}, S_{k+1}, X_{k+1}]$ depends on $T_k = \sum_{i=1}^k X_i$, which cannot be written as a function of $S_k, X_k, A_k, S_{k+1}, X_{k+1}$ only. In the proof of Lemma 1, we will demonstrate that this stronger conditional mean independence assumption is necessary, because it ensures the derivation of time-homogeneous properties of the value function.

2.3 Comparisons with existing frameworks

Assumptions (A1)-(A3) form a new framework for off-policy evaluation with irregularly-spaced, outcome-dependent observation times. For illustration, we provide comparisons of the proposed framework with existing approaches.

First, Figure 1 shows three data-generation schemes under the classic reinforcement learning framework (i.e. MDP framework), the SMDP framework and our proposed framework, respectively. It is not hard to see that, by assuming the observation times to be regularly spaced, the MDP framework in Figure 1(a) does not take into consideration gap times. The SMDP framework in Figure 1(b) introduces X_{k+1} to denote the stochastic sojourn

2.3 Comparisons with existing frameworks

time between states S_k and S_{k+1} , but the distribution of X_{k+1} is usually assumed to be fully determined by $\{S_k, A_k, S_{k+1}\}$. So, X_{k+1} is conditionally independent of X_k given the previous state S_k and action A_k . This restriction simplifies the derivation of the Bellman equations, but may be violated in practical applications. On the other hand, our proposed framework generalizes the SMDP in several aspects: (i) It relaxes the conditional independence assumption by allowing X_{k+1} to depend on the previous gap time X_k . (ii) It also allows the next state S_{k+1} to depend on the previous gap time X_k , providing a more flexible model for the state-action process. (iii) By allowing the action A_k to depend on both S_k and X_k , the proposed framework can evaluate a more general class of policies, and therefore may provide valuable insights into optimal policy learning.

Next, although the MDP framework can be extended by incorporating X_k into the state vector, existing policy evaluation methods can not be directly applied to estimate the value functions defined in Section 2.1. For example, by replacing S_k with $(S_k^\top, X_k)^\top$, the Bellman equation under the MDP framework (Shi et al., 2022) can be rewritten as

$$Q^\pi(S_k, X_k, A_k) = E\{\gamma R(T_{k+1}) + \gamma \sum_{a \in \mathbb{A}} Q^\pi(S_{k+1}, X_{k+1}, a) \pi(a|S_{k+1}, X_{k+1}) | S_k, X_k, A_k\}.$$

Although this equation accounts for the gap time X_k , we will demonstrate in Section 4 that this extension of the Bellman equation still results in

2.3 Comparisons with existing frameworks

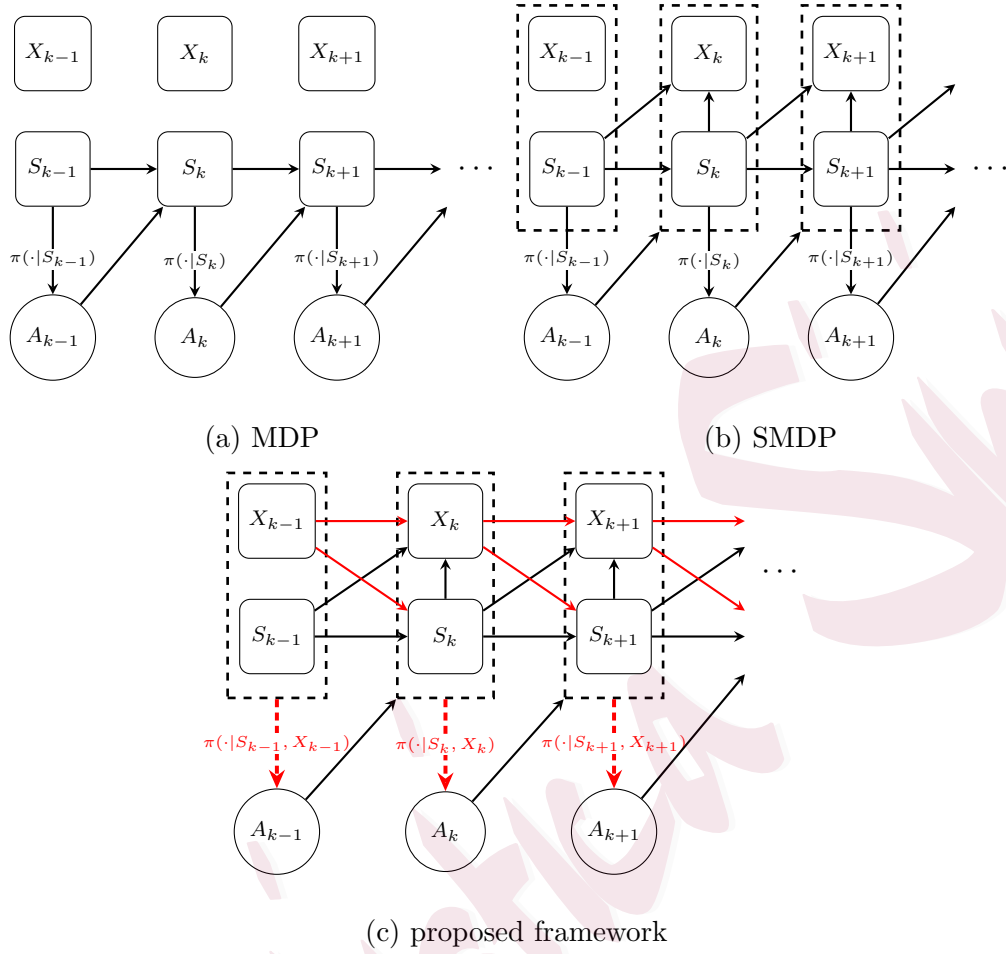


Figure 1: Data structures of MDP, SMDP, and the proposed framework, where the differences between the proposed framework and the SMDP framework are highlighted in red.

biased estimation for the value function under the proposed framework.

Lastly, we emphasize that the existing SMDP approaches also fail to provide unbiased estimates of the proposed value function, even when ex-

2.3 Comparisons with existing frameworks

panded state variables $(S_k^\top, X_k)^\top$ are used for constructing the policy. Specifically, the corresponding Bellman equation under the SMDP can be expressed as

$$Q^\pi(S_k, X_k, A_k) = E \left\{ \int_0^{X_{k+1}} \gamma^x R(T_k + x) dx + \gamma^{X_{k+1}} \sum_{a \in \mathbb{A}} Q^\pi(S_{k+1}, X_{k+1}, a) \right. \\ \left. \pi(a|S_{k+1}, X_{k+1}) \middle| S_k, X_k, A_k \right\}.$$

By incorporating the gap time X_{k+1} into both the discount term $\gamma^{X_{k+1}}$ and the integrated reward term $\int_0^{X_{k+1}} \gamma^x R(T_k + x) dx$, this formulation accounts for time-varying discounts, but also introduces challenges in computing the reward term, as the underlying reward process $R(t)$ is not always observable. The existing SMDP literature addresses this issue using different approximations. When $R(T_k + x)$ is explicitly known as a function of $\{x, S_k, S_{k+1}, A_k\}$, the integral can be evaluated directly from observed data (Crites and Barto, 1998). When $R(T_k + x)$ is unknown, the reward integral is commonly approximated by $\int_0^{X_{k+1}} \gamma^x dx R(T_{k+1})$ under the assumption that $R(t)$ remains constant over $(T_k, T_{k+1}]$ (Bradtke and Duff, 1994), or simply approximated by $\gamma^{X_{k+1}} R(T_{k+1})$, under the assumption that $R(T_{k+1})$ can be interpreted as a weighted reward accrued during $(T_k, T_{k+1}]$ (Barto and Mahadevan, 2003; Fatemi et al., 2022). It is not hard to see that all these approximations rely on rather strong assumptions about the reward

function and thus, as we will demonstrate in Section 4, may result in biased estimation of the target value function.

3. Off-Policy evaluation

3.1 Reformulation

Let n denote the number of trajectories in the observed dataset. For the i th trajectory, let $K_i \in \mathbb{Z}_+$ denote the number of observed decision stages, and let $\{S_{i,k}\}_{k \geq 0}$, $\{X_{i,k}\}_{k \geq 0}$, $\{A_{i,k}\}_{k \geq 0}$, and $\{R_i(T_{i,k+1})\}_{k \geq 0}$ denote the sequences of states, gap times, actions and rewards, respectively. Suppose the trajectories are independently generated according to a behavior policy $b(\cdot|\cdot, \cdot)$. Then, the observed dataset $\{X_{i,k}, S_{i,k}, A_{i,k}, R_i(T_{i,k+1}); k = 0, \dots, K_i, i = 1, \dots, n\}$ forms n i.i.d. copies of $\{S_k, X_k, A_k, R(T_{k+1})\}_{k \geq 0}$. Let $\pi(\cdot|\cdot, \cdot)$ denote a target policy which may differ from b . The goal of this section is to estimate the policy value of π based on the observed dataset generated by b .

We first restate assumption (A2) in an equivalent form using the notations of counting process theory. By viewing $\{X_k\}_{k \geq 1}$ as the gap times of a right-continuous counting process $\{N(t)\}_{t \geq 0}$ with jumps at $\{T_k\}_{k \geq 1}$, we introduce notations $T_{N(t)} = T_k$, $X_{N(t)} = X_k$, $A_{N(t)} = A_k$, $S_{N(t)} = S_k$, $S_{N(t)+1} = S_{k+1}$ for all $T_k \leq t < T_{k+1}$. The information collected over $[0, t)$

3.1 Reformulation

is denoted as $\mathcal{F}(t-) = \sigma \{ \{N(u)\}_{u < t}, \{S_{N(u)}\}_{u < t}, \{X_{N(u)}\}_{u < t}, \{A_{N(u)}\}_{u < t} \}$.

Then we can claim that, under (A1), the gap times $\{X_k\}_{k \geq 1}$ have the property defined in (A2) if and only if the corresponding observation process $\{N(t)\}_{t \geq 0}$ satisfies the following assumption.

(A2') The intensity function of the observation process $\{N(t)\}_{t \geq 0}$ conditional on $\mathcal{F}(t-)$ and $S_{N(t-)+1}$ satisfies

$$\begin{aligned} E \{dN(t) | \mathcal{F}(t-), S_{N(t-)+1}\} &= \lambda(t - T_{N(t-)}; S_{N(t-)+1}, S_{N(t-)}, X_{N(t-)}, A_{N(t-)}) dt \\ &= \lambda(t - T_k; S_{k+1}, S_k, X_k, A_k) dt, \end{aligned}$$

where $k = \max\{j; T_j < t\}$ and λ denotes the intensity function.

Here, (A2') assumes a modulated renewal process which is proposed by Cox (1972) for assessing dependencies in point processes. A detailed proof of the equivalence is presented in Section S3.1 of the Supplementary Material. For illustration, we provide two counterexamples to show the necessity of this modulated renewal process assumption. On one hand, if we relax the renewal process assumption and consider a general counting process with intensity $\lambda(t; S_{k+1}, S_k, X_k, A_k)$, then the hazard function of gap time X_{k+1} takes the form $\lambda(x + T_k; S_{k+1}, S_k, X_k, A_k)$, under which the time-homogeneity of $\{X_k\}_{k \geq 1}$ no longer holds due to the existence of T_k . On the other hand, if we extend the intensity model to a rate model, for

3.1 Reformulation

example, by introducing a state-dependent unobserved frailty $\{\xi_k\}_{k \geq 1}$, then the Markov assumption also fails since the distribution of X_{k+1} can depend on historical observations through $\{\xi_k\}_{k \geq 1}$ conditional on the current state, gap time, and action.

Now we present a reformulation of the value functions based on assumption (A2'). Without loss of generality, we assume the underlying reward process $R(t)$ is left-continuous, satisfying $R(t) \in \sigma\{\mathcal{F}(t-), S_{N(t-)+1}\}$. This indicates that $R(T_{k+1})$ is a function of $\{\{T_j, S_j, X_j, A_j\}_{j \leq k}, S_{k+1}, T_{k+1}\}$. Then together with (A2'), the Q-function $Q_{\mathcal{I},k}^\pi(S_k, X_k, A_k, T_k)$ can be rewritten as

$$\begin{aligned} & \mathbb{E}^\pi \left\{ \int_{T_k}^\infty \gamma^{u-T_k} R(u) du \middle| S_k, X_k, A_k, T_k \right\} \\ &= \mathbb{E}^\pi \left[\int_{T_k}^\infty \frac{\gamma^{u-T_k} R(u) E\{dN(u) | \mathcal{F}(u-), S_{N(u-)+1}\}}{\lambda(u - T_{N(u-)}; S_{N(u-)+1}, S_{N(u-)}, X_{N(u-)}, A_{N(u-)})} \middle| S_k, X_k, A_k, T_k \right] \\ &= \mathbb{E}^\pi \left\{ \int_{T_k}^\infty \frac{\gamma^{u-T_k} R(u) dN(u)}{\lambda(u - T_{N(u-)}; S_{N(u-)+1}, S_{N(u-)}, X_{N(u-)}, A_{N(u-)})} \middle| S_k, X_k, A_k, T_k \right\} \\ &= \mathbb{E}^\pi \left\{ \sum_{j \geq k+1} \frac{\gamma^{T_j-T_k} R(T_j)}{\lambda(X_j; S_j, S_{j-1}, X_{j-1}, A_{j-1})} \middle| S_k, X_k, A_k, T_k \right\} \\ &= \mathbb{E}^\pi \left\{ \sum_{j \geq k+1} \gamma^{T_j-T_k} R_{\mathcal{I}}(T_j) \middle| S_k, X_k, A_k, T_k \right\}, \end{aligned}$$

where $R_{\mathcal{I}}(T_j) = R(T_j)/\lambda(X_j; S_j, S_{j-1}, X_{j-1}, A_{j-1})$ is an inverse-intensity-weighted reward. This reformulation shows that the conditional expectation of the integrated reward can be viewed as the conditional expectation of a

3.2 Policy evaluation

sum of inverse-intensity-weighted rewards. Moreover, it is straightforward to verify that the inverse-intensity-weighted reward sequence $\{R_{\mathcal{I}}(T_k)\}_{k \geq 1}$ also satisfies assumption (A3). Using the above reformulation results, the following lemma further establishes the time-homogeneous properties of the integrated-reward-based value function $Q_{\mathcal{I},k}^{\pi}(s, x, a, t)$, with proofs provided in Section S3.2 of the Supplementary Material.

Lemma 1. *Under assumptions (A1)-(A3), the value function $Q_{\mathcal{I},k}^{\pi}(s, x, a, t)$ is free of k and t .*

Lemma 1 implies that, under the proposed framework, the value of a dynamic policy π is independent of both the number of past decision stages k and the observation initiation time t . Consequently, we can omit k and t , and denote the action-value function simply as $Q_{\mathcal{I}}^{\pi}(s, x, a)$ for brevity. The above result is essential for the classical MDP with fixed and regular decision time points. Here, we show that this is also true for the considered off-policy evaluation with irregularly-spaced, outcome-dependent observation times. Therefore, a Bellman equation can be similarly derived.

3.2 Policy evaluation

Now we develop estimates for $Q_{\mathcal{I}}^{\pi}(s, x, a)$, $V_{\mathcal{I}}^{\pi}(s, x)$ and $V_{\mathcal{I}}^{\pi}(\mathcal{G})$. Following the reinforcement learning techniques (Sutton and Barto, 2018), we first derive

3.2 Policy evaluation

a Bellman equation for $Q_{\mathcal{I}}^{\pi}(s, x, a)$, with a detailed derivation presented in Section S3.2 of the Supplementary Material.

Lemma 2. *Under assumptions (A1)-(A3), the action-value function $Q_{\mathcal{I}}^{\pi}(s, x, a)$ satisfies the Bellman equation*

$$Q_{\mathcal{I}}^{\pi}(S_k, X_k, A_k) = E\left\{ \int \gamma^{x'} \sum_{a \in \mathbb{A}} \pi(a|S_{k+1}, x') Q_{\mathcal{I}}^{\pi}(S_{k+1}, x', a) d\mathcal{P}_X(x'; S_{k+1}, S_k, X_k, A_k) \right. \\ \left. + \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) | S_k, X_k, A_k \right\}, \text{ for } k \geq 0.$$

The Bellman equation in Lemma 2 differs slightly from the standard version $Q_{\mathcal{I}}^{\pi}(S_k, X_k, A_k) = E\left\{ \gamma^{X_{k+1}} \sum_{a \in \mathbb{A}} \pi(a|S_{k+1}, X_{k+1}) Q_{\mathcal{I}}^{\pi}(S_{k+1}, X_{k+1}, a) + \gamma^{X_{k+1}} R_{\mathcal{I}}(T_{k+1}) | S_k, X_k, A_k \right\}$. This is because, following the discussion that (A2) can be reformulated as (A2'), we can construct an appropriate model for the gap times X_{k+1} . Then, the first term on the right-hand side of the standard Bellman equation, which is a function of X_{k+1} , can be replaced by its integration over X_{k+1} with respect to the conditional distribution function $\mathcal{P}_X$, thereby obtaining a distribution-modulated version of the Bellman equation to improve the estimation efficiency.

Next, we use the Bellman equation to construct estimating equations for $Q_{\mathcal{I}}^{\pi}(s, x, a)$. Consider a B-spline model which is indexed by $\theta \in \Theta$ and takes the form of $Q(s, x, a; \theta) = \xi(s, x, a)^{\top} \theta$, where $\xi(s, x, a) = \{\Phi_L^{\top}(s, x) \mathbb{I}(a = 0), \Phi_L^{\top}(s, x) \mathbb{I}(a = 1), \dots, \Phi_L^{\top}(s, x) \mathbb{I}(a = m - 1)\}^{\top}$, $\theta = (\theta_0^{\top}, \theta_1^{\top}, \dots, \theta_{m-1}^{\top})^{\top}$,

3.2 Policy evaluation

$\Phi_L(\cdot) = \{\Phi_{L,1}(\cdot, \cdot), \dots, \Phi_{L,L}(\cdot, \cdot)\}^\top$ is a vector of L B-spline basis functions, and θ_a is a L -dimensional parameter vector for $a = 0, \dots, m-1$. Assume $Q(s, x, a; \theta_\pi) = Q_{\mathcal{I}}^\pi(s, x, a)$ for some $\theta_\pi \in \Theta$. Then, following the Bellman equation, for $i = 1, \dots, n$ and $k = 0, \dots, K_i$, we have

$$E \left\{ \gamma^{X_{i,k+1}} R_{\mathcal{I},i}(T_{i,k+1}) + (\mathcal{U}_{\pi,i,k+1} - \xi_{i,k})^\top \theta_\pi \middle| S_{i,k}, X_{i,k}, A_{i,k} \right\} = 0,$$

where $\mathcal{U}_{\pi,i,k+1} = \int \gamma^{x'} \sum_a \pi(a|S_{i,k+1}, x') \xi(S_{i,k+1}, x', a) d\mathcal{P}_X(x'; S_{i,k+1}, S_{i,k}, X_{i,k}, A_{i,k})$ and $\xi_{i,k} = \xi(S_{i,k}, X_{i,k}, A_{i,k})$. Furthermore, by replacing $\lambda(\cdot)$ and $\mathcal{P}_X(\cdot)$ with their estimators $\hat{\lambda}(\cdot)$ and $\hat{\mathcal{P}}_X(\cdot)$, which will be detailed in the next subsection, we obtain the corresponding estimators of $R_{\mathcal{I},i}(T_{i,k+1})$ and $\mathcal{U}_{\pi,i,k+1}$. Denote them by $\hat{R}_{\mathcal{I},i}(T_{i,k+1})$ and $\hat{\mathcal{U}}_{\pi,i,k+1}$, respectively. Then, the estimating equation for θ can be

$$\sum_{i=1}^n \sum_{k=0}^{K_i-1} w(S_{i,k}, X_{i,k}, A_{i,k}) \left\{ \gamma^{X_{i,k+1}} \hat{R}_{\mathcal{I},i}(T_{i,k+1}) + (\hat{\mathcal{U}}_{\pi,i,k+1} - \xi_{i,k})^\top \theta \right\} = 0$$

for some predefined weighting function $w(s, x, a)$. For simplicity, we set $w(S_{i,k}, X_{i,k}, A_{i,k}) = \xi_{i,k}$. The resulting estimator of θ then takes the form

$$\hat{\theta}_\pi = \left\{ \sum_{i=1}^n \sum_{k=0}^{K_i-1} \xi_{i,k} (\xi_{i,k} - \hat{\mathcal{U}}_{\pi,i,k+1})^\top \right\}^{-1} \left\{ \sum_{i=1}^n \sum_{k=0}^{K_i-1} \xi_{i,k} \gamma^{X_{i,k+1}} \hat{R}_{\mathcal{I},i}(T_{i,k+1}) \right\},$$

and the value functions $Q_{\mathcal{I}}^\pi(s, x, a)$, $V_{\mathcal{I}}^\pi(s, x)$ and $V_{\mathcal{I}}^\pi(\mathcal{G})$ can be estimated by

$$\hat{Q}_{\mathcal{I}}^\pi(s, x, a) = \xi(s, x, a)^\top \hat{\theta}_\pi, \quad \hat{V}_{\mathcal{I}}^\pi(s, x) = \left\{ \sum_{a \in \mathbb{A}} \pi(a|s, x) \xi(s, x, a) \right\}^\top \hat{\theta}_\pi, \quad \text{and} \\ \hat{V}_{\mathcal{I}}^\pi(\mathcal{G}) = \left\{ \sum_{a \in \mathbb{A}} \int \pi(a|s, x) \xi(s, x, a) d\mathcal{G}(s, x) \right\}^\top \hat{\theta}_\pi, \text{ respectively.}$$

3.3 Modulated renewal process model

3.3 Modulated renewal process model

In this subsection, we develop estimators for λ and $\mathcal{P}_X$. As discussed in Section 3.1, the observation process $N(t)$ satisfying assumption (A2') forms a modulated renewal process. Therefore, as an illustration, we propose a Cox-type modulated renewal process model (Cox, 1972; Oakes and Cui, 1994; Lin et al., 2013), while alternative semiparametric models, such as the proportional odds model, accelerated failure time model, and additive model (Chen et al., 2018), can be adopted similarly. Specifically, assume that the intensity function satisfies

$$\lambda(t - T_k; S_{k+1}, S_k, X_k, A_k) dt = \lambda_0(t - T_k) \exp(\beta_0^\top Z_k) dt, \quad (3.1)$$

where $k = \max\{j; T_j < t\}$, $Z_k = \phi(S_{k+1}, S_k, X_k, A_k) \in \mathbb{R}^q$ for some pre-specified q -dimensional function ϕ , β_0 is a vector of unknown parameters and λ_0 is an unspecified baseline intensity function. Let $\{N_i(t), t \geq 0\}_{1 \leq i \leq n}$ denote the n independent and identically distributed trajectories of $N(t)$. In the i th trajectory, define $N_{i,k+1}(x) = N_i(T_{i,k} + x) - N_i(T_{i,k})$ and $Z_{i,k} = \phi(S_{i,k+1}, S_{i,k}, X_{i,k}, A_{i,k})$ for $k = 0, \dots, K_i - 1$. Let $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ denote the cumulative baseline function, and let τ be a positive constant satisfying $pr(X_{i,k} \geq \tau) > 0$. Following model (3.1), we obtain that $E\{dN_{i,k+1}(t) - I(X_{i,k+1} \geq t) \exp(\beta_0^\top Z_{i,k}) d\Lambda_0(t)\} = 0$ for all $i = 1, \dots, n$,

3.4 Asymptotic properties

$k = 0, \dots, K_i$ and $0 \leq t \leq \tau$. Thus, we construct estimating equations for the cumulative baseline hazard function $\Lambda_0(\cdot)$ and the regression coefficient β_0 as

$$\begin{aligned} \sum_{i=1}^n \sum_{k=0}^{K_i-1} \{dN_{i,k+1}(t) - I(X_{i,k+1} \geq t) \exp(\beta_0^\top Z_{i,k}) d\Lambda_0(t)\} &= 0, \quad 0 \leq t \leq \tau, \\ \int_0^\tau \sum_{i=1}^n \sum_{k=0}^{K_i-1} Z_{i,k} \{dN_{i,k+1}(t) - I(X_{i,k+1} \geq t) \exp(\beta_0^\top Z_{i,k}) d\Lambda_0(t)\} &= 0. \end{aligned}$$

Let $\hat{\beta}$ and $\hat{\Lambda}_0(\cdot)$ denote the solutions to the above estimating equations, i.e.,

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n \sum_{k=0}^{K_i-1} dN_{i,k+1}(t)}{\sum_{i=1}^n \sum_{k=0}^{K_i-1} I(X_{i,k+1} \geq t) \exp(\hat{\beta}^\top Z_{i,k})},$$

and $\hat{\beta}$ is the solution to $U(\beta) = 0$ with

$$U(\beta) = \sum_{i=1}^n \sum_{k=0}^{K_i-1} \int_0^\tau \left\{ Z_{i,k} - \frac{\sum_{j=1}^n \sum_{l=0}^{K_j-1} I(X_{j,l+1} \geq t) \exp(\beta^\top Z_{j,l}) Z_{j,l}}{\sum_{j=1}^n \sum_{l=0}^{K_j-1} I(X_{j,l+1} \geq t) \exp(\beta^\top Z_{j,l})} \right\} dN_{i,k+1}(t).$$

Let $K(\cdot)$ be a kernel function and b_n be the kernel bandwidth. Then we can estimate the intensity function $\lambda(x'; s', s, x, a)$ by $\hat{\lambda}(x'; s', s, x, a) = \hat{\lambda}_0(x') \exp(\hat{\beta}^\top z)$ with $\hat{\lambda}_0(x) = b_n^{-1} \int_0^\tau K\{b_n^{-1}(u - x)\} d\hat{\Lambda}_0(u)$, and estimate $\mathcal{P}_X$ by $\hat{\mathcal{P}}_X(x'; s', s, x, a) = 1 - \exp\{-\hat{\Lambda}_0(x') \exp(\hat{\beta}^\top z)\}$ for $z = \phi(s', s, x, a)$.

3.4 Asymptotic properties

Now, we present the asymptotic properties of the estimated policy value $\hat{V}_T^\pi(\mathcal{G})$. In the following theorem, assumption (A4) is a smoothness condition commonly used in nonparametric estimation to control the approx-

3.4 Asymptotic properties

imation power of sieve bases, assumptions (A5) and (A6) ensure the existence of $\widehat{\theta}_\pi$, and assumption (A7) is required for constructing asymptotic approximation results when $K_i \rightarrow \infty$. (C1)-(C5) are regularity conditions commonly adopted in the studies of modulated renewal processes (Pons and de Turckheim, 1988; Lin et al., 2013), and (C6) ensures the existence of $\sigma_{\pi, \mathcal{G}}^{-1}$ by assuming $\lambda_{\min}(\Omega_\pi) > 0$. For brevity, we relegate these conditions, together with the proof of the theorem, to Section S3.3 of the Supplementary Material.

Theorem 1. *Suppose assumptions (A1)-(A7), conditions (C1)-(C6), and model (3.1) hold. Denote $N_K = \sum_{i=1}^n K_i$. Suppose $L \ll N_K^{1/2}/\log(N_K)$, $L^{2p/(d+1)} \gg N_K(1 + \|\zeta_{\pi, \mathcal{G}}\|_2^{-1})^2$, $N_K b_n^4 \rightarrow 0$ and $N_K b_n^2 \rightarrow \infty$, and there exists some constant $c_R \geq 0$ such that $\text{pr}(\max_k |R(T_k)| \leq c_R) = 1$. Then as $N_K \rightarrow \infty$, we have $N_K^{1/2} \sigma_{\pi, \mathcal{G}}^{-1} \{\widehat{V}_\pi^\pi(\mathcal{G}) - V_\pi^\pi(\mathcal{G})\} \xrightarrow{d} N(0, 1)$. Here, $D_\pi = E\{N_K^{-1} \sum_{i=1}^n \sum_{k=0}^{K_i-1} \xi_{i,k}(\xi_{i,k} - \mathcal{U}_{\pi, i, k+1})^\top\}$, $\sigma_{\pi, \mathcal{G}}^2 = \zeta_{\pi, \mathcal{G}}^\top D_\pi^{-1} \Omega_\pi (D_\pi^\top)^{-1} \zeta_{\pi, \mathcal{G}}$, $\zeta_{\pi, \mathcal{G}} = \sum_{a \in \mathbb{A}} \int \pi(a|s, x) \xi(s, x, a) d\mathcal{G}(s, x)$, and Ω_π is given in Section S3.3 of the Supplementary Material.*

Here, Theorem 1 provides a bidirectional asymptotic property for the policy value functions. Whether n is fixed with $K_i \rightarrow \infty$ or K_i remains bounded with $n \rightarrow \infty$, we always have $N_K = \sum_{i=1}^n K_i \rightarrow \infty$, and hence the weak convergence holds. The corresponding property for the estimated

state-value function $\widehat{V}^\pi(s, x)$ is also bidirectional, which follows directly by setting $\mathcal{G}$ as a Dirac delta function at a fixed point $(s, x) \in \mathbb{S} \times \mathbb{R}^+$.

4. Simulation

In this section, we conduct simulation studies to examine the finite-sample performance of the proposed estimates. We begin by describing our setup. Firstly, for the initial values, we generate S_0 from a uniform distribution on $[-1.5, 1.5]$, generate X_0 from an exponential distribution with intensity 0.5, and generate A_0 according to a behavior policy $b(\cdot; S_0, X_0)$, where $b(a; s, x)$ can be one of the three distributions: (b1) Bernoulli distribution with mean 0.5; (b2) Bernoulli distribution with mean $\text{sigmoid}(s)$; (b3) Bernoulli distribution with mean $0.5\text{sigmoid}(s) + 0.5\text{sigmoid}(x)$. Furthermore, given the current state S_k , gap time X_k , and action A_k , the next state S_{k+1} is generated from one of the three models: (S1) $S_{k+1} = 0.75(2A_k - 1)S_k + \epsilon_k$; (S2) $S_{k+1} = \{0.75 - 0.25I(X_k < 0.5)\}(2A_k - 1)S_k + \epsilon_k$; (S3) $S_{k+1} = \{0.75 - 0.25I(X_k < 0.5) + 0.25I(X_{k+1} > 1)\}(2A_k - 1)S_k + \epsilon_k$, with $\{\epsilon_k\}_{k \geq 1} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.25)$. The next gap time X_{k+1} is generated from one of the following hazard functions: (X1) $\lambda(x) = \lambda_0(x) \exp(0.5X_k)$; (X2) $\lambda(x) = \lambda_0(x) \exp(-S_k + 0.5X_k + A_k - 0.5S_kA_k)$; (X3) $\lambda(x) = \lambda_0(x) \exp(-S_k + 0.5X_k + A_k - 0.5S_kA_k + 0.5S_{k+1})$, where the baseline function $\lambda_0(\cdot)$ takes

one of the following forms: $\lambda_0(x) \equiv 1.5$, $\lambda_0(x) = 0.5x$, and $\lambda_0(x) = 0.5x^2$.

The reward of action A_k , which is observed at T_{k+1} , is defined as $R(T_{k+1}) = (S_{k+1} - S_k - A_k + 0.5)X_{k+1} + r_k$ with $\{r_k\}_{k \geq 0} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.25)$.

From the myriad combinations of generative models, we select 20 representative scenarios for a brief illustration, which can be further classified into four data generation schemes. Detailed settings of the four schemes are listed in Table S1, and plots for the corresponding data structures are illustrated in Figure S1 of the Supplementary Material. In general, all four schemes satisfy assumptions (A1)-(A3), but vary in the dependence structures among states, actions, and gap times. In Scheme 1, the generation of X_{k+1} does not depend on the values of (S_k, A_k) , and the value of X_k does not affect the generation of (S_{k+1}, A_{k+1}) either. This represents the scenarios where the observation process $N(t)$ is independent of the state-action transition process. In Scheme 2, the generation of gap times is allowed to be correlated with the actions and states, while X_{k+1} and S_{k+1} are conditionally independent of each other given (S_k, X_k, A_k) . More complicated settings that accommodate correlation between X_{k+1} and S_{k+1} are given in Schemes 3 and 4, with Scheme 3 conforming to the observation process model (3.1), whereas Scheme 4 departs from the model (3.1), thereby providing an assessment of the robustness of the estimators developed in

Section 3.3.

In the policy evaluation procedure, the target policy π is defined as a member of the following class of linear policies $\pi(a; s, x) = aI(\alpha_0 + \alpha_1 s + \alpha_2 x > 0) + (1 - a)I(\alpha_0 + \alpha_1 s + \alpha_2 x \leq 0)$, and two values of the coefficient $(\alpha_0, \alpha_1, \alpha_2)$, i.e., $(0, -1, 0)$ and $(-1, -1, 1)$, are considered for each generated dataset. To calculate the true value of $V_{\mathcal{I}}^{\pi}(\mathcal{G})$, we simulate $N = 6 \times 10^5$ independent long trajectories where (S_0, X_0) is distributed according to $\mathcal{G}$, and $\{A_k\}_{k \geq 0}$ are chosen according to the target policy π . Then, $V_{\mathcal{I}}^{\pi}(\mathcal{G})$ is approximated by $N^{-1} \sum_{j=1}^N \sum_{l=0}^{K_j-1} \gamma^{T_{j,l+1}} R_j(T_{j,l+1}) / \lambda(X_{j,l+1}; S_{j,l+1}, S_{j,l}, X_{j,l}, A_{j,l})$. Moreover, since state S_k and gap time X_k may not have bounded supports, we transform the data by $S_k^* = \Phi(S_k)$ and $X_k^* = 1 - \exp(-X_k)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable, and $1 - \exp(-x)$ is the cumulative distribution function of an exponential random variable with rate 1. The basis functions are constructed from the tensor product of two cubic B-spline sets, whose knots are placed at equally spaced sample quantiles of the transformed states and gap times. We set $L = \lfloor N_K^{3/14} \rfloor$. When estimating the B-spline coefficients θ , three methods were considered, namely, the proposed method, the SMDP method, and the MDP method. Specifically, the proposed estimator $\hat{\theta}_{\pi}$ is constructed from the distribution-modulated Bellman equation in Lemma 2, whereas

the SMDP and MDP estimators are obtained by solving the Bellman equations given in Section 2.3, which treat the gap time X_k as a part of the state variable.

The performances of the three estimation methods are summarized in Tables S2-S4 of the Supplementary Material. From the results, we observe that the proposed method performs well under Schemes 1-3. The estimators are asymptotically unbiased, the estimated standard errors are close to the standard deviations, and the coverage probabilities of the 95% confidence intervals approach the nominal level when either n or K increases. In comparison, the MDP and SMDP estimators are biased across all schemes, including Scheme 1, where gap times $\{X_k\}_{k \geq 1}$ are independent of the states $\{S_k\}_{k \geq 1}$. This implies that simply extending MDP and SMDP models by incorporating gap times into the state variable will introduce bias into policy evaluation, even when the observation process is independent of the state-action process empirically.

Lastly, since the proposed policy evaluation method requires modeling the observation process $N(t)$, we also investigate the robustness of the proposed estimates. Specifically, we generate data following Scheme 4, which violates model (3.1), and calculate the estimates using the proposed methods developed under model (3.1). Thus, all estimates are obtained under

a misspecified model. From the results reported in Table S5 of the Supplementary Material, the proposed estimates demonstrate limited bias and provide satisfactory variance estimation. Hence, the proposed methods are robust to mild misspecification of the observation process model (3.1).

5. Application: HealthPartners Data

In this section, we illustrate our proposed method by applying it to the periodontal recall interval selection problem from a HealthPartners database of longitudinal electronic health records tracking periodontal disease status and progression, along with other covariates, among subjects in the Minneapolis, Minnesota area (Guan et al., 2020). While the current insurance-mandated standard of care sets a recall interval of 6 months for all patients, in practice, recall frequencies are closely related to patient outcomes, provider workload, and dental healthcare costs, making the length of recall intervals an ongoing topic of debate and research (Riley et al., 2013). Recently, NICE guidelines (Clarkson et al., 2018) suggested that recall intervals should vary over time and be determined based on disease levels and oral disease risk. However, the selection of an appropriate personalized recall interval is a multifaceted clinical decision that is challenging to evaluate mechanistically.

The goal of our study is to evaluate recall interval recommendation policies based on the patients' dental visit records. Specifically, our HealthPartners dataset consists of 7654 dental visit records from 1056 patients enrolled within the HealthPartners electronic health records from January 1, 2007, to December 31, 2014. For each visit record, we have recall visit date, probing pocket depth (PPD, measured in mm) from all available tooth sites as the periodontal disease assessment endpoint, and the clinician-recommended recall interval for the follow-up visit. The total follow-up time for these subjects has a median of 5.95 years, with a maximum of 7.97 years (approximately 8 years of data). The frequency of recall visits has a median of 8, with a range from 3 to 29. The gap time (actual recall intervals) has a median of 181 days (approximately 6 months), with a range of 1 to 2598 days. Due to the unavailability of full-mouth, site-level measures of clinical attachment level (another important biomarker for periodontal disease) in the HealthPartners database, we refrain from using the ACES framework (Holtfreter et al., 2024), the modified 2018 periodontal status classification scheme for epidemiological survey data, and instead consider the mean PPD (mean across all tooth sites at a specific visit) as a plausible subject-level periodontal disease endpoint, with a median of 2.51 mm (range: 0.0-6.2 mm).

To apply our methods, let T_k denote the record time, let X_k denote the gap time between T_k and T_{k-1} , let S_k denote (mean) PPD measurements at T_k , and let A_k be an indicator of the recommended recall interval at T_k , where $A_k = 1$ implies the patient is advised to revisit within 6 months, and $A_k = 0$ otherwise. The discount factor is set as $\gamma = 0.9$. The record time T_k and gap time X_k are rescaled by dividing the actual time in days by 180, so that the median of X_k is 1. Within the HealthPartners database, the mean PPD measurements (S_k) are significantly correlated (p -value < 0.001) with the lengths of clinician-recommended recall intervals, suggesting that the mean PPD measurements play a crucial role in the dentist's decisions regarding revisit dates. On the other hand, although other covariates are available in the dataset, only PPD and clinical attachment loss (CAL) show a significant correlation with the treatment decisions made by dentists. Given the high correlation between PPD and CAL, we select PPD as the only state variable. For the reward function, since $\text{PPD} \leq 3$ mm is typically considered as a threshold to distinguish periodontally healthy sites from diseased ones (Ireland, 2010), we define $R(t) = I\{S(t) \leq 3\} + I\{S(t) > 3\}(4 - S(t))$. Based on this definition, $R(\cdot)$ represents a continuous function where $R(t)$ denotes the score of the potential PPD measurement at time t . Therefore, the value function $V_{\mathcal{I}}^{\pi}(s, x)$, which integrates the patient's

PPD measurements over the entire treatment process, provides a reasonable evaluation of the treatment policies.

In the policy evaluation procedure, we considered three target policies for comparison, namely, the observed stochastic policy π_{os} , the observed fixed policy π_{of} , and the optimal policy π_{op} . Detailed definitions of the target policies are provided in Section S2 of the Supplementary Material, and plots of the obtained policies are shown in Figure S2. Typically, under the observed policies π_{of} and π_{os} , patients with deeper pocket depth and more frequent recalls are more likely to be recommended a shorter recall interval, while under the estimated optimal policy π_{op} , the relationship among actions, states, and gap times takes a more complicated nonlinear form.

The estimated policy value functions are summarized in Figure 2. In general, the estimated values change with not only the initial state but also the initial gap time, which indicates that the frequency of recall visits does affect the performance of the policies and should be taken into account in the evaluation process. Moreover, by comparing the estimated results obtained from the proposed method, MDP method, and SMDP method, we observe that the results from the MDP and SMDP methods exhibit significant differences from the proposed one in both the magnitudes of

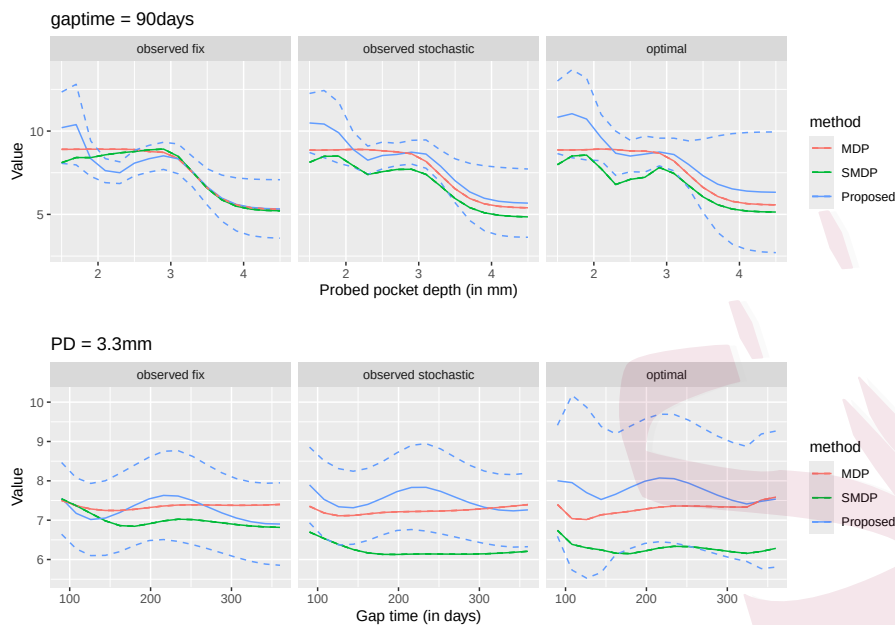


Figure 2: Plots of the estimated value (solid line) and confidence intervals (dashed lines) of the observed fixed policy π_{of} , the observed stochastic policy π_{os} , and the estimated optimal policy π_{op} with gap time fixed at 90 days (upper panel) and probing pocket depth fixed at 3.3 mm (lower panel), which are obtained by the MDP method (red line), SMDP method (green line), and the proposed method (blue line).

values and the patterns of the function.

Lastly, we compare the estimated value functions of the observed stochastic policy, observed fixed policy, and the optimal policy under the modulated method. Plots of the corresponding confidence intervals are provided in Fig-

ure S3 of the Supplementary Material. From the results, it is evident that the changes in action assignments made by the optimal policy, compared to the other two observed policies, have enhanced the policy value.

6. Discussion

In this paper, we study the off-policy evaluation problem for infinite-horizon dynamic policies with irregularly spaced and outcome-dependent observation times. We first introduce a policy evaluation criterion based on integrated reward, rather than the classic cumulative-reward value function, to better align with practical settings with continuous observation times. To ensure the estimability of the target value functions, we construct a new framework which incorporates not only the state-action process, but also an observation process characterizing the irregularly spaced decision times. The proposed framework is also more flexible than the SMDP framework by allowing the gap times between observation times to be correlated with previous states, actions, and gap times, as well as to affect future states and actions. This, in turn, enables a more comprehensive utilization of the information hidden in the observation times, and thus leads to more reasonable evaluations for dynamic policies.

In the estimation procedure, we used an inverse-intensity-weighted re-

ward $R_{\mathcal{I},i}(T_{i,k+1}) = R_i(T_{i,k+1})/\lambda(X_{i,k+1})$ to construct estimating equations for the target value functions. To mitigate the numerical instability caused by small values in the denominator, it would be advantageous to construct another null intensity model (Lin et al., 2004) and incorporate its estimated intensity into the weight function of the estimating equation. In addition, if the policy used to collect data is available, we can further incorporate the action-selection probabilities into the estimation procedure to construct doubly robust estimators of policy value, and obtain their efficiency bounds (Kallus and Uehara, 2022). Finally, although our paper focuses on the off-policy evaluation of dynamic policies, it is not hard to adapt the proposed method to search for an optimal dynamic policy (as shown in Section S2 of the Supplementary Material), as well as to make inferences on the value of the estimated optimal dynamic policy.

Supplementary material

The supplementary material includes the proofs of the results in Section 3.1, Lemmas 1-2 and Theorem 1, as well as additional numerical results from the simulation and the application, respectively.

REFERENCES

Acknowledgement

The authors gratefully acknowledge funding support from the National Natural Science Foundation of China, Grant Number 12001371; National Science Foundation, Grant Number DMS-2113637 and SES 2242776; United States National Institutes of Health, Grant Number R01DE031134.

References

- Azar, M. G., V. Gómez, and H. J. Kappen (2012). Dynamic policy programming. *The Journal of Machine Learning Research* 13, 3207–3245.
- Barto, A. G. and S. Mahadevan (2003, Oct). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems* 13(4), 341–379.
- Bradtke, S. and M. Duff (1994). Reinforcement learning methods for continuous-time markov decision problems. *Advances in Neural Information Processing Systems* 7, 393–400.
- Chen, X., J. Ding, and L. Sun (2018). A semiparametric additive rate model for a modulated renewal process. *Lifetime Data Analysis* 24, 675–698.
- Clarkson, J. E., N. B. Pitts, D. Bonetti, D. Boyers, H. Braid, R. Elford, P. A. Fee, R. Floate, B. Goulão, G. Humphris, et al. (2018). INTERVAL dental recalls trial: a multicentre randomised controlled trial investigating the best dental recall interval for optimum, cost-effective maintenance of oral health in dentate adults attending dental primary care. *BMC Oral Health* 18, 1–10.

REFERENCES

- Clarkson, J. E., N. B. Pitts, B. Goulao, D. Boyers, C. R. Ramsay, R. Floate, H. J. Braid, P. A. Fee, F. S. Ord, H. V. Worthington, et al. (2020). Risk-based, 6-monthly and 24-monthly dental check-ups for adults: the INTERVAL three-arm RCT. *Health Technology Assessment (Winchester, England)* 24(60), 1–138.
- Cox, D. R. (1972). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes.*, pp. 55–66. New York: Wiley.
- Crites, R. H. and A. G. Barto (1998, Nov). Elevator group control using multiple reinforcement learning agents. *Machine Learning* 33(2), 235–262.
- Das, T. K., A. Gosavi, S. Mahadevan, and N. Marchallick (1999). Solving semi-markov decision problems using average reward reinforcement learning. *Management Science* 45(4), 560–574.
- Du, J., J. Futoma, and F. Doshi-Velez (2020). Model-based reinforcement learning for semi-markov decision processes with neural odes. *Advances in Neural Information Processing Systems* 33, 19805–19816.
- Fatemi, M., M. Wu, J. Petch, W. Nelson, S. J. Connolly, A. Benz, A. Carnicelli, and M. Ghassemi (2022). Semi-markov offline reinforcement learning for healthcare. In *Proceedings of the Conference on Health, Inference, and Learning*, Volume 174, pp. 119–137.
- Guan, Q., B. J. Reich, E. B. Laber, and D. Bandyopadhyay (2020). Bayesian nonparametric policy search with application to periodontal recall intervals. *Journal of the American Statistical Association* 115, 1066–1078.

REFERENCES

- Holtfreter, B., K. Kuhr, K. Borof, M. S. Tonetti, M. Sanz, K. Kornman, S. Jepsen, G. Aarabi, H. Völzke, T. Kocher, et al. (2024). ACES: A new framework for the application of the 2018 periodontal status classification scheme to epidemiological survey data. *Journal of Clinical Periodontology* 51, 512–521.
- Ireland, R. (2010). *A Dictionary of Dentistry* (1 ed.). Oxford, U.K.: Oxford University Press.
- Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48, New York, USA, pp. 652–661. PMLR.
- Kallus, N. and M. Uehara (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research* 21, 1–63.
- Kallus, N. and M. Uehara (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research* 70(6), 3282–3302.
- Lin, F.-C., Y. K. Truong, and J. P. Fine (2013). Robust analysis of semiparametric renewal process models. *Biometrika* 100(3), 709–726.
- Lin, H., D. O. Scharfstein, and R. A. Rosenheck (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66(3), 791–813.
- Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems* 31, 5356–5366.

REFERENCES

- Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2019). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association* 115, 692–706.
- Maei, H. R. and R. S. Sutton (2010). Gq(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the 3d Conference on Artificial General Intelligence (2010)*, pp. 100–105. Atlantis Press.
- Murphy, S. A. (2003). Optimal Dynamic Treatment Regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65, 331–355.
- Oakes, D. and L. Cui (1994). On semiparametric inference for modulated renewal processes. *Biometrika* 81(1), 83–90.
- Parr, R. E. (1998). *Hierarchical Control and Learning for Markov Decision Processes*. Ph. D. thesis, University of California, Berkeley.
- Pons, O. and E. de Turckheim (1988). Cox’s periodic regression model. *The Annals of Statistics* 16, 678–693.
- Riley, P., H. V. Worthington, J. E. Clarkson, and P. V. Beirne (2013). Recall intervals for oral health in primary care patients. *Cochrane Database of Systematic Reviews* 12, 1–54.
- Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, New York, NY, pp. 189–326. Springer.

REFERENCES

- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association* 100(469), 322–331.
- Shi, C., S. Luo, Y. Le, H. Zhu, and R. Song (2022). Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *Journal of the American Statistical Association* 119(545), 232–245.
- Shi, C., M. Uehara, J. Huang, and N. Jiang (2022). A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, Volume 162, pp. 20057–20094. PMLR.
- Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical Inference of the Value Function for Reinforcement Learning in Infinite-Horizon Settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84, 765–793.
- Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction*. Cambridge: MIT press.
- Tang, X., Z. T. Qin, F. Zhang, Z. Wang, Z. Xu, Y. Ma, H. Zhu, and J. Ye (2019). A deep value-network based approach for multi-driver order dispatching. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.
- Uehara, M., C. Shi, and N. Kallus (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 1–27.

REFERENCES

Yang, S. (2021). Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations. *Biometrics* 78, 937–949.

Yang, S., K. Pieper, and F. Cools (2019). Semiparametric Estimation of Structural Failure Time Models in Continuous-Time Processes. *Biometrika* 107, 123–136.

School of Statistics and Mathematics; Interdisciplinary Research Institute of Data Science,
Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China. E-mail: chenxin@amss.ac.cn

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695,
U.S.A. E-mail: wlu4@ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695,
U.S.A. E-mail: syang24@ncsu.edu

Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia 23220,
U.S.A. E-mail: dbandyop@vcu.edu